EERINGMATHEMATICALENGINEERINGMATHEMA

E. Ostertag

Mono- and Multivariable Control and Estimation

Linear, Quadratic and LMI Methods

Extra Materials
extras.springer.com

Springer

Mono- and Multivariable Control and Estimation

Mathematical Engineering

For further volumes:
http://www.springer.com/series/8445

Eric Ostertag

Mono- and Multivariable Control and Estimation

Linear, Quadratic and LMI Methods

Springer

Professor Emeritus Eric Ostertag
University of Strasbourg
LSIIT Laboratory
Control Systems, Vision and Robotics
bd Sébastien Brant, BP 10413
F-67412 Illkirch Cedex, France
Eric.Ostertag@unistra.fr

Additional material to this book can be downloaded from
http://extra.springer.com/2010/978-3-642-13733-4

ISBN 978-3-642-13733-4 e-ISBN 978-3-642-13734-1
DOI 10.1007/978-3-642-13734-1
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2010936361

Cover design: WMXDesign GmbH, Heidelberg

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

In the field of control systems, one area which has been studied extensively in the academic environment this past decade has encountered however as far as now relatively low acceptance in the industrial world. It is the area of control of multiple input plants, or of estimation of the states of multiple output plants. One exception is the decoupling control, but here also merely one or two design methods have found real applications. On another hand, though or maybe even because of the importance of digital control in practical realizations, it is essential to be at ease with these control or estimation methods, both in continuous and in discrete time, so as to be able to simulate them on a computer in the two forms. It is common practice, indeed, to describe an industrial process or plant initially by a continuous model, then to design and test adequate control schemes in continuous time, before switching to discrete time for the final computer implementation.

The book essentially consists of three parts. The first part, composed of Chapters 1 and 2, deals with the concepts of linear control laws and observers. A relatively unknown generalizing formula for multivariable controller or observer design, which encompasses some of the other methods while removing certain of their limitations, is also given. The basic prerequisites for the reader are here ground knowledge in Matrix Calculus and Linear Algebra. Concise reviews are given in the Appendices on these topics and the chapters are written with a clear mathematical approach, with proofs, so that the student as well as the engineer can follow the developments completely, without unnecessary details. The second part of the book, covering Chapters 3 to 5, deals with quadratic methods to determine control or estimation laws. The basics required to follow these chapters are essentially the same as before, except for Chapter 4 which requires some knowledge of Probability Calculus, the main concepts of which are also summarized in an Appendix. The third part of the book is represented by a single chapter which describes a rather new design method, valid as well for controllers as for observers, relatively unknown to the best of our knowledge from the industrial world: the method based on Linear Matrix Inequalities. This approach is presented here in a simple but rigorous way, so that no special prerequisite is needed to understand it. The reader might be surprised to discover how easily problems of the linear part as well as of the quadratic part can be translated into this elegant formalism, and how easily they can then be solved by available, freely downloadable software.

In all chapters, the continuous-time and discrete-time situations are treated almost always in parallel. The description tool used for the plants and for the mathematical developments is essentially the state space representation, sometimes also called internal representation, which is particularly suitable for multivariable systems. Many new methods for such plants are simply not tractable with other representations. Their use in this book with single input or single output systems, to which they apply evidently also, serves essentially to establish the basic theorems and their proofs in a highly readable way, before generalizing to the multivariable case.

Many exercises are given at the end of each chapter, with their complete solutions, most of them obtained in the MATLAB® environment. The programs used for their solving are grouped in an archive, which is downloadable from either the Springer or the MathWorks Internet site. The user can modify them at will by adding other solutions or other plant models of his own. These exercises represent in our mind the most important part of the learning process, and the reader is strongly encouraged to try the solutions by himself before reading them in the book. The simulations of these exercises have been realized by three Simulink® diagrams, given in one of the Appendices and also included in the downloadable archive for the reader's use.

This book is aimed at several kinds of public. It addresses first students in Engineering Schools or Universities, at the Master's level. It is also aimed at Control Engineers, faced with the control of industrial processes, where new specifications reaching far beyond the possibilities of the classical PID controller are imposed. Finally, of course, this book is also written for researchers wanting to broaden their knowledge of control approaches, in particular in the multivariable case.

Among the persons who contributed to the present shape of this book, the author is grateful to Maria Joana Carvalho, Assistant Professor at the Ecole Nationale Supérieure de Physique of the University of Strasbourg, for her helpful comments and suggestions, and to Gilles Duc, Professor at Supélec (formerly Ecole Supérieure d'Electricité), for his careful reading and valuable suggestions of improvement in several parts of the book. I would like to express here also my gratitude to Günter Roppenecker, Professor at the University Erlangen-Nürnberg, for his kind support in clarifying the design method bearing his name.

The author is also deeply indebted to Eva Hestermann-Beyerle, Senior Engineering Editor at Springer, for her kind invitation to publish this book in this new Springer series and helpful advices.

Suggestions of improvements and corrections about the book itself or the accompanying software will be of course greatly appreciated and can be sent to Eric.Ostertag@unistra.fr.

Summer 2010 Eric Ostertag

Contents

NOTATIONAL CONVENTIONS

- Scalar variables and scalar functions are denoted by lowercase Latin and Greek letters.

  Examples: $y(t)$ ε

  Exception: T_s, sampling time of discrete-time signals

- For discrete-time functions, the sample instant is indicated as a subscript, rather than in parentheses, in order to simplify the notations.

  Examples: y_{k+1}, and not $y(k+1)$

- Vectors are denoted by lowercase bold Latin and Greek letters.

  Examples: $\mathbf{u}_k$ $\boldsymbol{\gamma}$

- Matrices are denoted by capital bold Latin and Greek letters.

  Examples: $\mathbf{A}$ $\boldsymbol{\Phi}$

- Transforms (Laplace, z) of scalar or vectorial functions use the corresponding uppercase Latin and Greek letters:

Time function	Transform
$x(t)$, $\mathbf{x}(t)$	$\mathcal{L}[x(t)] = X(s)$, $\mathcal{L}[\mathbf{x}(t)] = \mathbf{X}(s)$
y_k, $\mathbf{y}_k$	$\mathcal{L}[y_k] = Y(z)$, $\mathcal{L}[\mathbf{y}_k] = \mathbf{Y}(z)$

- Transposition symbol: superscript T

  Examples: $\mathbf{A}^{\mathrm{T}}$ $\mathbf{c}^{\mathrm{T}}$

- Special notations for stochastic environments:
  - scalar or vectorial random variables and scalar or vectorial random functions are denoted by capital Latin letters.

    Examples: $X(t)$, Y_k, $\mathbf{V}(t)$, $\mathbf{W}_k$
  - deterministic realizations of such random variables or functions are denoted by the corresponding lowercase letter.

    Examples: $x(t)$, y_k, $\mathbf{v}(t)$, $\mathbf{w}_k$

N.B.: there should be no confusion with the above convention concerning the use of capital letters, for Laplace or z transforms and for matrices. In the first case, the dependent variable, s or z, will always be given explicitly. In the second case, no confusion should arise, since no *random matrices* will appear in this book.

- Estimates symbol: the upper hat symbol ($\widehat{\ }$)

  Examples: $\widehat{\mathbf{x}}$ $\widehat{\mathbf{x}}(t)$ $\widehat{\mathbf{x}}_{k|k}$

- Estimation errors symbol: the upper tilde symbol ($\tilde{\ }$)

  Examples: $\widetilde{\mathbf{x}}$ $\widetilde{\mathbf{x}}(t)$ $\widetilde{\mathbf{x}}_k$

  This symbol is also used, without risk of ambiguity, for matrices and vectors of *augmented* systems:

  Examples: $\widetilde{\mathbf{A}}$ $\widetilde{\mathbf{L}}$

- Similitude transformations, or basis changes: matrix and vectors resulting from such changes are denoted by a circumflex accent.

  Examples: $\hat{\mathbf{A}}$, $\hat{\mathbf{v}}$

- Linear systems:

  - state vector: $\mathbf{x}(t)$ or $\mathbf{x}_k \in \mathbb{R}^n$ (real vectors, of dimension *n*)
  - input: $\mathbf{u}(t)$ or $\mathbf{u}_k \in \mathbb{R}^p$ (real vectors, of dimension *p*)
  - output: $\mathbf{y}(t)$ or $\mathbf{y}_k \in \mathbb{R}^q$ (real vectors, of dimension *q*)

1 State Space Control Design

For all questions relating to the definitions and properties of the state-space representations of linear time-invariant (LTI) systems, and to the notations used in this book, the reader is invited to consult Appendix A at the end of the book as well as the page on Conventional Notations at its beginning.

A deliberate choice has been made not to describe in this book how these state space models are established, nor to tackle the numerous other modelization or model reduction methods existing for such systems. The reader may consult for this purpose various text books, such as [AlSa04], [Lev96], [Oga02], [ZhDG96].

To underline better the resemblance of many methods and formulae used in the continuous case and in the discrete case, this chapter and the following ones will deal simultaneously with both system types. The particulars of each type (formulae or proofs, e.g.) will be mentioned and developed, if necessary, in parallel. The two types of systems are distinguished all along this book by the different notations involved in their state representations.

Care has been taken also to identify differently in their notations the *monovariable* systems, including systems with only one input or one output or both, the so-called *single input – single output* (SISO) systems, and the *multivariable* systems, commonly designated in the control literature by *multiple input – multiple output* (MIMO) systems. These two abbreviations will be used throughout this manual.

1.1 State Space Control: an Introduction

1.1.1 Problem Description

We shall suppose in all the following that the process to be controlled, hereafter called the *plant* or the *open-loop system*, is a linear time-invariant system. Furthermore, to simplify the notations, only systems for which $\mathbf{D} = 0$ will be considered here, which is the great majority of real systems (see Sect. A.1.1 and A.1.2).

The plant will thus be described by one of the following state representations:

E. Ostertag, *Mono- and Multivariable Control and Estimation,*
DOI 10.1007/978-3-642-13734-1_1, © Springer-Verlag Berlin Heidelberg 2011

Continuous Case: $\begin{cases}\dot{\mathbf{x}}(t)=\mathbf{A}\,\mathbf{x}(t)+\mathbf{B}\,\mathbf{u}(t)\\ \mathbf{y}(t)=\mathbf{C}\,\mathbf{x}(t)\end{cases}$ or: *Discrete Case:* $\begin{cases}\mathbf{x}_{k+1}=\mathbf{\Phi}\,\mathbf{x}_k+\mathbf{\Gamma}\mathbf{u}_k\\ \mathbf{y}_k\quad=\mathbf{C}\,\mathbf{x}_k\end{cases}$

with: **A** or **Φ** : *system* matrix, constant, real, of size $(n\times n)$, i.e. $\in\mathbb{R}^{n\times n}$;

x: *state* vector, real, of dimension *n*, i.e. $\in\mathbb{R}^{n}$;

u: *input* vector, real, of dimension $p\le n$, i.e. $\in\mathbb{R}^{p}$;

B or **Γ** : *input* matrix, constant, real, of size $(n\times p)$, i.e. $\in\mathbb{R}^{n\times p}$;

y: *output* or *measurement* vector, real, of dimension $q\le n$, i.e. $\in\mathbb{R}^{q}$;

C: *output* or *measurement* matrix, constant, real, of size $(q\times n)$, $\in\mathbb{R}^{q\times n}$.

The system being at initial time t_0 $(=0$ for time-invariant systems) in the initial state $\mathbf{x}_0$, the aim is to transfer it to the desired final state, or *operating* state, $\mathbf{x}_f$, and this, while submitting the transient phase to given dynamic requirements, such as rise time or transient damping. Once the operating state $\mathbf{x}_f$ is reached, the output quantity **y** must usually be equal to an imposed reference value $\mathbf{y}_r$.

1.1.2 Solution: State Feedback

The way the problem is solved consists in feeding back the state vector **x** to the input by means of a feedback matrix **L**, as shown in Fig. 1.1 in the continuous case. **x** is supposed here to be fully accessible, a restriction which will be reconsidered in Sect. 1.1.5 (Remark 1.3). Such a control law is called *state feedback.*

1.1.3 Continuous Case

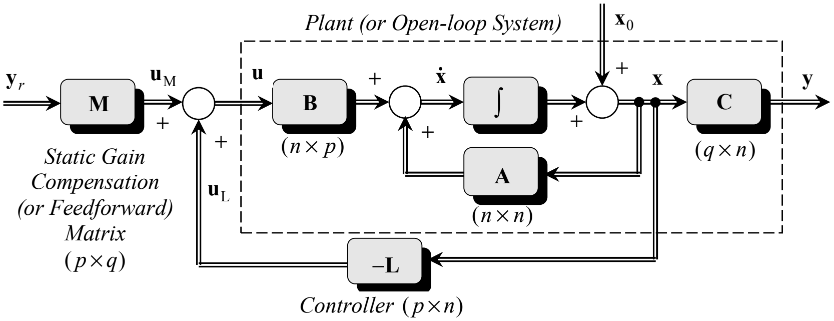


Fig. 1.1 State feedback control of a continuous time plant.

L is a constant $(p \times n)$ matrix called *state controller*. Its practical realization will consist of a set of constant elements, thus of proportional terms. It appears thus to play the role of a proportional controller, having in the extreme case of n^2 elements. Note however that the quantities calculated this way are proportional not only to the output but also to all its derivatives, as shown e.g. by (A.1) where $\dot{x}_i = x_{i+1}$. The state-feedback controller is thus rather equivalent to a PD controller. However it does not contain any "I"-term (integrator), point which will be addressed in Sect. 1.8. The role of the **M** matrix will be explained in Sect. 1.1.3.2.

The equations of the system shown in Fig. 1.1 are the following:

$$\dot{\mathbf{x}} = \mathbf{A}\,\mathbf{x} + \mathbf{B}\,\mathbf{u}\,;\;\; \mathbf{y} = \mathbf{C}\,\mathbf{x}\,; \tag{1.1}$$

$$\mathbf{u} = \mathbf{u}_{\mathrm{L}} + \mathbf{u}_{\mathrm{M}} = -\mathbf{L}\,\mathbf{x} + \mathbf{M}\,\mathbf{y}_r \;. \tag{1.2}$$

Note that the time dependency of the time-varying elements of these equations has been omitted. This will be the case throughout this book for continuous-time systems. The above equations lead to the following closed-loop state equation:

$$\dot{\mathbf{x}} = (\mathbf{A} - \mathbf{BL})\,\mathbf{x} + \mathbf{B}\,\mathbf{M}\,\mathbf{y}_r \;. \tag{1.3}$$

The closed-loop system is thus described by a state equation, with system matrix

$$\mathbf{A}_{CL} = \mathbf{A} - \mathbf{BL}\,, \tag{1.4}$$

and input matrix

$$\mathbf{B}_{CL} = \mathbf{BM}\,. \tag{1.5}$$

Since two types of external quantities act on this control system, the initial state $\mathbf{x}_0$ and the reference signal $\mathbf{y}_r$, two kinds of behavior can be considered.

1.1.3.1 Regulation Behavior

The regulation methods, which will be described in the following sections, aim at transferring the system state vector from the initial state $\mathbf{x}_0$ to the operating (final) state $\mathbf{x}_f$ under given conditions. This will result from an adequate choice of **L**.

The system being linear, it is then possible, in order to calculate **L**, to set $\mathbf{y}_r = 0$, according to the linear superposition principle. Equation (1.3) simplifies then to $\dot{\mathbf{x}} = (\mathbf{A} - \mathbf{BL})\,\mathbf{x}$. Once the equilibrium is reached (steady state), $\dot{\mathbf{x}} = 0$. It results necessarily in this case, provided all closed-loop eigenvalues have been chosen in the left-half *s* plane, thus that $\mathbf{A} - \mathbf{BL}$ is invertible, that

$$\mathbf{x}_f = 0\,. \tag{1.6}$$

1.1.3.2 Servo Behavior: Choice of Gain Compensation (or Feedforward) Matrix

In the steady state, which corresponds to $\dot{\mathbf{x}} = 0$, the equality $\mathbf{y} = \mathbf{y}_r$ must hold. In this state, (1.3) is written

$$0 = (\mathbf{A} - \mathbf{BL})\mathbf{x} + \mathbf{B}\mathbf{M}\mathbf{y}_r\,,$$

therefore:

$$\mathbf{x} = (\mathbf{BL} - \mathbf{A})^{-1}\mathbf{B}\mathbf{M}\mathbf{y}_r\,.$$

The output equation yields then

$$\mathbf{y} = \mathbf{y}_r = \mathbf{C}(\mathbf{BL} - \mathbf{A})^{-1}\mathbf{B}\mathbf{M}\mathbf{y}_r\,,$$

from where follows that

$$\mathbf{C}(\mathbf{BL} - \mathbf{A})^{-1}\mathbf{B}\mathbf{M} = \mathbf{I}\,. \tag{1.7}$$

Case where p = q: this case is often encountered in practice, where one has generally to associate with each output signal a reference signal, which is expected to influence as wanted this output. Such plants are called *square* plants. **M** is then a square, thus possibly invertible, matrix, and (1.7) yields

$$\boxed{\mathbf{M} = \left[\mathbf{C}(\mathbf{BL} - \mathbf{A})^{-1}\mathbf{B}\right]^{-1}}\,. \tag{1.8}$$

Remark 1.1. Not surprisingly, the **M** matrix is nothing else than the inverse of the closed-loop static gain. This gain is obtained, indeed, by letting $s = 0$ in the expression of the transfer matrix of a system in state-space representation, namely, for the closed-loop system with the assumption $\mathbf{D} = 0$ (see Appendix A, Sect. A.5), $\mathbf{G}_{CL}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A}_{CL})^{-1}\mathbf{B}$:

$$\mathbf{G}_{CL}(s)\,\big|_{s=0} = \mathbf{C}\left[s\mathbf{I} - (\mathbf{A} - \mathbf{BL})\right]^{-1}\mathbf{B}\,\Big|_{s=0} = \mathbf{C}(\mathbf{BL} - \mathbf{A})^{-1}\mathbf{B}\,. \tag{1.9}$$

Remark 1.2. The static gain compensation realized by the **M** matrix is ideal only if the plant parameters, on which the value of **M** depends, are perfectly known and do not vary with time. To compensate for these two restrictions, it may be advantageous to add an integral term to the control loop, as will be seen in Sect. 1.8.

Case where $p \neq q$, most often $p < q$: since the matrix in the second member of (1.9) is no longer square, thus not invertible, **M** cannot be calculated any more by (1.8). It is then possible to apply the unity static gain requirement only between a number of outputs equal to that of available control inputs and their corresponding reference values. This amounts to choosing a submatrix of **C** which has only *p* rows, which brings back to the previous case $(p = q)$. This is the purpose of the selection matrix $\mathbf{S}_c$ introduced in the simulation diagrams of Appendix D. For systems with $p > q$ one will have, on the contrary, to drop $p - q$ control inputs of the plant for that requirement.

1.1.4 Discrete Case

Let us represent this time the plant in its general form by a state-variable flow graph, where T_s is the sampling time (see Fig. 1.2):

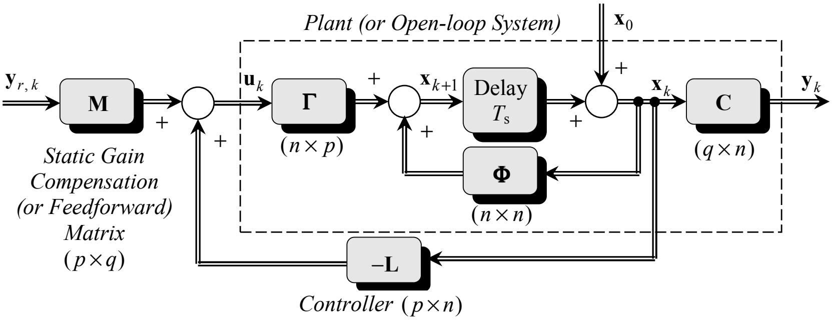


Fig. 1.2 State feedback control of a discrete-time plant.

The corresponding equations are here

$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k \,, \tag{1.10}$$

$$\mathbf{x}_{k+1} = \mathbf{\Phi}\,\mathbf{x}_k + \mathbf{\Gamma}\mathbf{u}_k \,, \tag{1.11}$$

$$\mathbf{u}_k = -\mathbf{L}\mathbf{x}_k + \mathbf{M}\mathbf{y}_{r,k} \ . \tag{1.12}$$

$\mathbf{x}_k$ is assumed here again totally accessible. The closed-loop state equation is derived by substituting (1.12) into (1.11):

$$\mathbf{x}_{k+1} = (\mathbf{\Phi} - \mathbf{\Gamma}\mathbf{L})\mathbf{x}_k + \mathbf{\Gamma}\mathbf{M}\mathbf{y}_{r,k} \ . \tag{1.13}$$

The closed-loop system is again described by a state equation, with system matrix

$$\mathbf{\Phi}_{CL} = \mathbf{\Phi} - \mathbf{\Gamma L}, \tag{1.14}$$

and input matrix:

$$\mathbf{\Gamma}_{CL} = \mathbf{\Gamma M}. \tag{1.15}$$

1.1.4.1 Regulation Behavior

As in the continuous case, let us set $\mathbf{y}_{r,k} = 0, \forall k$. The design problem will then amount to determine a sequence of control values $\mathbf{u}_0, \mathbf{u}_1, \cdots, \mathbf{u}_k$, obtained from $\mathbf{x}_k$ by $\mathbf{u}_k = -\mathbf{L}\mathbf{x}_k$, which drives the system from an arbitrary initial state $\mathbf{x}_0$ to the origin ($\mathbf{x}_f = 0$).

1.1.4.2 Servo Behavior

In the steady state, which corresponds here to $\mathbf{x}_{k+1} = \mathbf{x}_k$, it is still required to have $\mathbf{y}_k = \mathbf{y}_{r,k}$. Equation (1.13) becomes

$$\mathbf{x}_k = (\mathbf{\Phi} - \mathbf{\Gamma L})\mathbf{x}_k + \mathbf{\Gamma M}\mathbf{y}_{r,k}.$$

Therefore

$$\mathbf{x}_k = (\mathbf{I} + \mathbf{\Gamma L} - \mathbf{\Phi})^{-1}\mathbf{\Gamma M}\mathbf{y}_{r,k},$$

and

$$\mathbf{y}_k = \mathbf{C}(\mathbf{I} + \mathbf{\Gamma L} - \mathbf{\Phi})^{-1}\mathbf{\Gamma M}\mathbf{y}_{r,k} = \mathbf{y}_{r,k},$$

yielding

$$\mathbf{C}(\mathbf{I} + \mathbf{\Gamma L} - \mathbf{\Phi})^{-1}\mathbf{\Gamma M} = \mathbf{I},$$

and, if $p = q$:

$$\mathbf{M} = \left[\mathbf{C}(\mathbf{I} + \mathbf{\Gamma L} - \mathbf{\Phi})^{-1}\mathbf{\Gamma}\right]^{-1}. \tag{1.16}$$

N.B.: the $\mathbf{M}$ matrix is here again equal to the inverse of the static gain of the closed-loop system, gain which follows from the formula giving the transfer matrix of a discrete-time system of system matrix $\mathbf{\Phi}$ and direct feedthrough matrix $\mathbf{D} = 0$, namely $\mathbf{G}(z) = \mathbf{C}(z\mathbf{I} - \mathbf{\Phi})^{-1}\mathbf{\Gamma}$, applied to the closed-loop system with the substitution $z = 1$:

$$\mathbf{G}_{CL}(z)\big|_{z=1} = \mathbf{C}\left[z\mathbf{I} - (\mathbf{\Phi} - \mathbf{\Gamma}\mathbf{L})\right]^{-1}\mathbf{\Gamma}\Big|_{z=1} = \mathbf{C}\left[\mathbf{I} + \mathbf{\Gamma}\mathbf{L} - \mathbf{\Phi}\right]^{-1}\mathbf{\Gamma}\,.$$

1.1.5 Design Methods of the Feedback Matrix L

In the following sections, two large classes of design methods of the **L** matrix will be described**.**

- The so-called *Pole Placement Method*, or *Pole Assignment Method,* where the closed-loop poles are given in advance (they are *placed* in the *s* or *z* plane) and where **L** is determined such that the closed loop gets really these poles. Strictly speaking, the values imposed are the *eigenvalues* of the continuous or discrete system matrix of the closed-loop system. The language abuse stems from the fact that the poles of the closed-loop system are contained among its eigenvalues (Appendix A, Sect. A.6).

  This method is well adapted to continuous or discrete systems of SISO or MIMO type, but it does not make any use of the additional degree of freedom available for the latter type, as will be seen later: as a matter of fact it does not use any of the possibilities offered by the *eigenstructure* of the closed loop system.

  The *Finite Settling-Time Design* [Kal60], [MuBa64], represents a particular case of the previous design method, exclusively reserved to the discrete systems. It consists in placing all the poles of the closed-loop discrete transfer function at the origin of the complex plane (*z*-plane). It is the transcription in the state-variable domain of the design method of a minimum settling-time RST-controller, [AsWi97] [GoOs03].

  As a matter of fact, this method goes far beyond the scope of this frame, because it applies also to nonlinear systems, for which one cannot speak anymore of pole placement. Incidentally, the approach for the derivation of the corresponding algorithm, to be described in Sect. 1.3, will depart entirely from the pole placement technique.

- The design methods specific to MIMO systems, which still assign the eigenvalues of the closed-loop system, but address also its structure, more precisely its *eigenvectors* or the *decoupling* of certain quantities.

  - The *Simple Modal Control* or *Modal Decoupling* [Ros62] is a pole placement in diagonal representation, but limited to a number of poles not exceeding the number of process inputs.
  - The *Input-Output Decoupling Method* permits in some cases to put the closed-loop system in the form of several SISO systems connected in parallel.

- The *Complete Modal Control* extends the modal control approach to the whole set of closed-loop eigenvalues, by introducing *invariant parameter vectors*, and encompasses as special cases the two previous methods.
- Finally, a *General Formula* for the synthesis of both MIMO controllers and observers is presented, which removes some of the restrictions of the complete modal design.

Remark 1.3. It was assumed until now that the state vector $\mathbf{x}$ was completely accessible. In fact, the design will occur in two steps:

1. $\mathbf{x}$ is assumed accessible, and the state controller $\mathbf{L}$ of the *control law* $\mathbf{u} = -\mathbf{L}\mathbf{x}$ is designed according to one of the mentioned methods;
2. if $\mathbf{x}$ is not fully accessible or not accessible at all, one will then proceed with the design of a *state reconstructor*, the purpose of which will be to yield the *best possible estimation* $\hat{\mathbf{x}}(t)$, respectively $\hat{\mathbf{x}}_k$, of the state vector from what is available to measurement, i.e. from the measurement (or output) vector $\mathbf{y}(t)$, respectively $\mathbf{y}_k$, and the control vector $\mathbf{u}(t)$, respectively $\mathbf{u}_k$.

In the pure deterministic situation, such a state reconstructor will be called an *observer*, e.g. a *Luenberger observer*, the synthesis of which will be addressed in Chapter 2 of this book.

In the stochastic situation, i.e. in the presence of noisy signals, such a state reconstructor will be an *estimating filter* or a *predictive filter*, e.g. a *Kalman filter*, the most famous of this type of reconstructors, the synthesis of which will be addressed in Chapter 4 of this book, entitled *Optimal Linear Filtering*.

Remark 1.4. There are other approaches for the design of a state-feedback control law, totally different from the ones mentioned here above.

One of them consists in looking for a control law which minimizes an algebraic criterion, or cost function, usually quadratic, without any concern about the resulting places of the closed-loop poles: it is called *Optimal Control*. Due to its importance and its distinctive characteristics, this design method will be the subject of a separate chapter, Chapter 3 of this book. The duality existing between the formalism of that approach and the one used in optimal filtering will be presented at the end of Chapter 4.

One more method is the *Generalized Predictive Control*, or *GPC* [ClMo87], [BoDu96], where it is searched to minimize on a finite horizon a cost function, which involves both the future, or predicted, response of the process to present and future control signals, supposed known, and the resulting control signals. We will not go into this subject in the frame of this book.

Still another class of methods is that of *Robust Control*, among which should be cited the most known, the H_∞ optimization method. They will not either be the subject of this book. Robustness of optimal control loops and its loss due to the inclusion of an estimator into the loop will nevertheless be discussed in Chap. 3 and 5, respectively.

1.2 Pole Placement Method

1.2.1 Basics of the Method

Consider the system in Fig. 1.3. Since now only the regulation behavior is of concern, thus the reaction of the system to the initial condition $\mathbf{x}_0$, we can let $\mathbf{y}_r = 0$. Furthermore, the output equation $\mathbf{y} = \mathbf{C}\mathbf{x}$ will be ignored, since it has already been taken into account at the time of designing the gain compensation term **M**. We deal here only with the *regulation of the state vector* **x**.

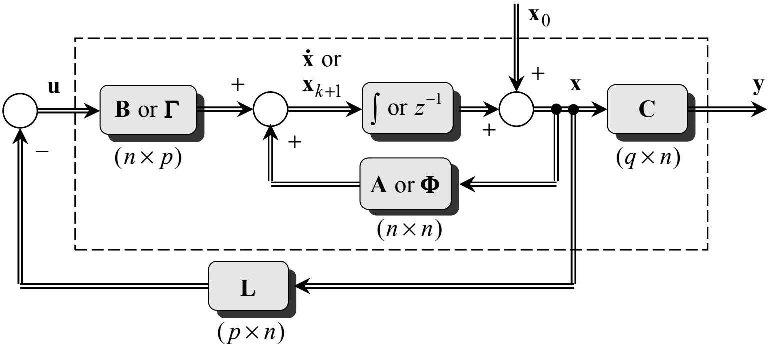


Fig. 1.3 Regulation diagram of a state feedback controlled system

The design by pole placement method consists in choosing the constant state-feedback matrix **L** such that the eigenvalues $\lambda_{L1}, \lambda_{L2}, \cdots, \lambda_{Ln}$ of the closed-loop system, represented above in the continuous and the discrete cases, are situated at given locations in the complex *s* or *z* plane. Since, most of the time, these eigenvalues are identical to the closed-loop poles, as already recalled previously, the term *poles* will often be used for them by language abuse.

Conditions of existence of the matrix L. The question arises now about the existence of such a matrix **L**. [Won67] has shown that, for *all* the eigenvalues of a system to be displaceable to new arbitrary values $\lambda_{L1}, \lambda_{L2}, \cdots, \lambda_{Ln}$ by means of state feedback, it is necessary and sufficient that this system be (completely) controllable. For the *mathematical* controllability, or, in case of uncertain or bad conditioned plant matrices, *practical* controllability criterions, see Appendix A.

So what happens, if the plant is not *completely* controllable? If this is the case, it is possible, by means of a change of state variables $\mathbf{x} = \mathbf{T}\mathbf{z}$, to transform it to the controllability canonical decomposition (Sect. A.2.3, Appendix A), where $\mathbf{z}_C$ ($\mathbf{z}_{\bar{C}}$) are the controllable (uncontrollable) parts of the new state vector **z**:

$$\begin{pmatrix} \dot{\mathbf{z}}_C \\ \dot{\mathbf{z}}_{\bar{C}} \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{A}}_C & \hat{\mathbf{A}}_{12} \\ 0 & \hat{\mathbf{A}}_{\bar{C}} \end{pmatrix} \begin{pmatrix} \mathbf{z}_C \\ \mathbf{z}_{\bar{C}} \end{pmatrix} + \begin{pmatrix} \hat{\mathbf{B}}_C \\ 0 \end{pmatrix} \mathbf{u} .$$

Applying now a state feedback control as given by (1.2), i.e.,

$$\mathbf{u} = -\begin{pmatrix} \mathbf{L}_1 & \mathbf{L}_2 \end{pmatrix} \begin{pmatrix} \hat{\mathbf{z}}_C \\ \hat{\mathbf{z}}_{\bar{C}} \end{pmatrix} + \mathbf{M}\,\mathbf{y}_r ,$$

yields the following closed loop system:

$$\begin{pmatrix} \dot{\mathbf{z}}_C \\ \dot{\mathbf{z}}_{\bar{C}} \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{A}}_C - \hat{\mathbf{B}}_C\mathbf{L}_1 & \hat{\mathbf{A}}_{12} - \hat{\mathbf{B}}_C\mathbf{L}_2 \\ 0 & \hat{\mathbf{A}}_{\bar{C}} \end{pmatrix} \begin{pmatrix} \mathbf{z}_C \\ \mathbf{z}_{\bar{C}} \end{pmatrix} + \begin{pmatrix} \hat{\mathbf{B}}_C\mathbf{M} \\ 0 \end{pmatrix} \mathbf{y}_r .$$

The closed-loop system matrix being block-diagonal, its eigenvalues are the union of the eigenvalues of $\hat{\mathbf{A}}_C - \hat{\mathbf{B}}_C\mathbf{L}_1$ and those of $\hat{\mathbf{A}}_{\bar{C}}$. The choice of the closed-loop eigenvalues is thus not completely arbitrary, since it must contain the eigenvalues of $\hat{\mathbf{A}}_{\bar{C}}$. These eigenvalues represent the dynamic behavior of the uncontrollable part of the plant, which is not modified by the feedback.

Choice of the poles:

This choice is guided by practical considerations. If a second order type behavior is desired for the closed loop, one could choose for instance one dominant complex conjugated pole pair plus one real pole of multiplicity order $(n-2)$, where *n* is the system order. The choice of one unique real pole of multiplicity order *n* is of course also possible.

The exact location of these poles will be deduced from the dynamic behavior desired for the controlled system. The closed-loop state equation is, according to (1.3) with $\mathbf{y}_r = 0$:

$$\dot{\mathbf{x}} = (\mathbf{A} - \mathbf{BL})\,\mathbf{x} . \tag{1.17}$$

The eigenvalues $\lambda_{L1}, \lambda_{L2}, \cdots, \lambda_{Ln}$ of the closed-loop system matrix $\mathbf{A} - \mathbf{BL}$, which are now supposed chosen, are the zeros of the closed-loop characteristic polynomial

$$\det\left[s\,\mathbf{I} - (\mathbf{A} - \mathbf{BL})\right] = s^n + \alpha_{n-1}(\mathbf{L}) \cdot s^{n-1} + \cdots + \alpha_0(\mathbf{L}) , \tag{1.18}$$

where the notation $\alpha_i(\mathbf{L})$ is used simply to recall that the coefficients of this polynomial depend on the elements of the unknown matrix **L**. Choosing arbitrarily these eigenvalues amounts to choosing arbitrarily the characteristic polynomial coefficients, since

$$p(s) = (s - \lambda_{L1})(s - \lambda_{L2}) \cdots (s - \lambda_{Ln}) = s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0 \,.$$

By coefficient identification, the following equations are obtained:

$$\begin{aligned} \alpha_0(\mathbf{L}) &= p_0 \\ \alpha_1(\mathbf{L}) &= p_1 \\ &\vdots \\ \alpha_{n-1}(\mathbf{L}) &= p_{n-1} \end{aligned} \,, \tag{1.19}$$

i.e. a set of n equations in $p \times n$ unknowns, the elements of the **L** matrix.

Consequences:

- the solution is determined if $p = 1$ (single-input systems);
- it is undetermined (underdetermined set) if $p > 1$ (multiple-input systems).

1.2.2 Single-input Systems

In this case $(p = 1)$, the **L** matrix is reduced to one unique row[1],

$$\mathbf{L} = \boldsymbol{\ell}^{\mathrm{T}} = \begin{pmatrix} \ell_1 & \ell_2 & \cdots & \ell_n \end{pmatrix}, \tag{1.20}$$

and the control law, which is scalar here, becomes according to (1.2):

$$u = -\boldsymbol{\ell}^{\mathrm{T}} \mathbf{x} \,. \tag{1.21}$$

As long as n does not exceed 3 or 4, or if the closed-loop system matrix is sufficiently empty, the system of equations (1.19) can be solved directly to get $\boldsymbol{\ell}^{\mathrm{T}}$.

If on the contrary the plant to be controlled is of a higher order, it is better to use the following design method, which will be first described hereafter in the case of a plant in controllability canonical representation, and transposed then to the case of an arbitrary representation.

[1] By exception with the typographic convention adopted in this book, the symbol $\boldsymbol{\ell}^{\mathrm{T}}$ will be used here instead of $\mathbf{l}^{\mathrm{T}}$, in order to avoid confusion with the identity matrix $\mathbf{I}$.

1.2.2.1 Plants in Controllability Canonical Form

If the plant is given in controllability canonical form, the closed-loop system matrix is, according to (1.3) or (1.17), joined to (1.20),

$$\mathbf{A}_{CL} = \mathbf{A}_C - \mathbf{b}_C \boldsymbol{\ell}_C^{\mathrm{T}}, \tag{1.22}$$

where $\mathbf{A}_C$ and $\mathbf{b}_C$ are given by (A.1) or (A.31) of Appendix A, recalled here:

$$\mathbf{A}_C = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 \\ -a_0 & -a_1 & \cdots & -a_{n-1} \end{pmatrix}, \quad \mathbf{b}_C = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

This yields

$$\mathbf{A}_{CL} = \begin{pmatrix} 0 & 1 & 0 \cdots & 0 \\ 0 & 0 & 1 \ddots & \vdots \\ \vdots & & \ddots \ddots & 0 \\ \vdots & & \ddots & 1 \\ -a_0 & \cdots & \cdots \cdots & -a_{n-1} \end{pmatrix} - \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} \ell_1 & \ell_2 & \cdots & \ell_n \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \cdots & 0 \\ 0 & 0 & 1 \ddots & \vdots \\ \vdots & & \ddots \ddots & 0 \\ \vdots & & & 1 \\ -a_0 - \ell_1 & \cdots & \cdots \cdots & -a_{n-1} - \ell_n \end{pmatrix}.$$

As can be seen, the obtained matrix has still the controllability canonical form. The coefficients of the closed-loop characteristic polynomial are thus obtained without calculation, since in that form the last row of the system matrix, here $\mathbf{A}_{CL}$, is composed of the characteristic equation coefficients, except the one of the highest power term, with change of sign, according to Rule 1 of Sect. A.7.2 of Appendix A. Thus the following theorem:

Theorem 1.1. *Suppose that, in the control system of Fig. 1.3, the plant is continuous-time, single-input, and is given in controllability canonical form, with the characteristic polynomial*

$$s^n + a_{n-1} s^{n-1} + \cdots + a_1 s + a_0,$$

and that, for the closed-loop system, the characteristic polynomial

$$s^n + p_{n-1} s^{n-1} + \cdots + p_1 s + p_0$$

is prescribed. Then, the **L** *matrix which will implement it is given by*

$$\mathbf{L} = \boldsymbol{\ell}_C^{\mathrm{T}} = (p_0 - a_0 \quad p_1 - a_1 \quad \cdots \quad p_{n-1} - a_{n-1}) \ . \tag{1.23}$$

1.2.2.2 Plants in Arbitrary State-representation (Ackermann's Formula)

If the plant is controllable, it is always possible to transform it to the controllability canonical form by means of the following coordinate transformation (see Sect. A.7.2 of Appendix A):

$$\mathbf{z} = \mathbf{V}\,\mathbf{x} \,,$$

where $\mathbf{z}$ is the state vector characterizing this canonical form, and where the inverse transform matrix, $\mathbf{V} = \mathbf{T}^{-1}$, is given by the following formula (equation (A.32) of Appendix A):

$$\mathbf{V} = \begin{pmatrix} \mathbf{q}_C^{\mathrm{T}} \\ \mathbf{q}_C^{\mathrm{T}}\,\mathbf{A} \\ \vdots \\ \mathbf{q}_C^{\mathrm{T}}\,\mathbf{A}^{n-1} \end{pmatrix},$$

where $\mathbf{q}_C^{\mathrm{T}}$ is the last row of the controllability matrix inverse, thus of $\mathbf{Q}_C^{-1}$.

In the controllability canonical form, the control law is written, according to (1.21), as

$$u = -\boldsymbol{\ell}_C^{\mathrm{T}}\,\mathbf{z} \,,$$

which yields

$$u = -\boldsymbol{\ell}_C^{\mathrm{T}}\,\mathbf{V}\,\mathbf{x} \ .$$

With the use of (1.23), the feedback matrix $\boldsymbol{\ell}_C^{\mathrm{T}}$ has then, in an arbitrary state representation, the following expression:

$$\boldsymbol{\ell}^{\mathrm{T}} = \boldsymbol{\ell}_C^{\mathrm{T}}\,\mathbf{V} = \left(p_0 - a_0 \quad p_1 - a_1 \quad \cdots \quad p_{n-1} - a_{n-1}\right) \begin{pmatrix} \mathbf{q}_C^{\mathrm{T}} \\ \mathbf{q}_C^{\mathrm{T}}\,\mathbf{A} \\ \vdots \\ \mathbf{q}_C^{\mathrm{T}}\,\mathbf{A}^{n-1} \end{pmatrix}$$

$$= (p_0 - a_0)\,\mathbf{q}_C^{\mathrm{T}} + (p_1 - a_1)\,\mathbf{q}_C^{\mathrm{T}}\mathbf{A} + \cdots + (p_{n-1} - a_{n-1})\,\mathbf{q}_C^{\mathrm{T}}\mathbf{A}^{n-1} \ . \tag{1.24}$$

To continue this calculation, a well known theorem from matrix analysis will be used.

Cayley-Hamilton Theorem. *Every square matrix* $\mathbf{A}$*, of size* $(n \times n)$*, satisfies its own characteristic equation. In other words, if the characteristic equation of* $\mathbf{A}$ *is*

$$p(\lambda) = \det(\lambda \mathbf{I} - \mathbf{A}) = \lambda^n + \alpha_{n-1}\lambda^{n-1} + \cdots + \alpha_1 \lambda + \alpha_0 = 0\,,$$

then the following holds:

$$p(\mathbf{A}) = \mathbf{A}^n + \alpha_{n-1}\,\mathbf{A}^{n-1} + \cdots + \alpha_1\,\mathbf{A} + \alpha_0\,\mathbf{I} = 0\ . \tag{1.25}$$

It is known moreover that two similar matrices have the same eigenvalues. This is thus the case of $\mathbf{A}$ and $\mathbf{A}_C = \mathbf{V}\,\mathbf{A}\,\mathbf{V}^{-1}$ (see Sect. A.7.1 and Equation (A.31) of Appendix A). They will thus have also the same characteristic polynomial. According to the Cayley-Hamilton theorem, $\mathbf{A}$ must thus satisfy the characteristic equation of $\mathbf{A}_C$,

$$\mathbf{A}^n + \alpha_{n-1}\,\mathbf{A}^{n-1} + \cdots + \alpha_1\,\mathbf{A} + \alpha_0\,\mathbf{I} = 0\,,$$

yielding

$$-a_0\,\mathbf{I} - a_1\,\mathbf{A} - \cdots - a_{n-1}\,\mathbf{A}^{n-1} = \mathbf{A}^n\,,$$

and, by left multiplication by $\mathbf{q}_C^{\mathrm{T}}$,

$$-a_0\,\mathbf{q}_C^{\mathrm{T}} - a_1\,\mathbf{q}_C^{\mathrm{T}}\mathbf{A} - \cdots - a_{n-1}\,\mathbf{q}_C^{\mathrm{T}}\,\mathbf{A}^{n-1} = \mathbf{q}_C^{\mathrm{T}}\,\mathbf{A}^n\ . \tag{1.26}$$

Substituting (1.26) in (1.24) yields then the remarkable following result [Ack72]:

Theorem 1.2. *If the plant* $\dot{\mathbf{x}} = \mathbf{A}\,\mathbf{x} + \mathbf{b}\,u$ *of a single-input control system is controllable, and if the closed-loop characteristic polynomial*

$$p(s) = s^n + p_{n-1}\,s^{n-1} + \cdots + p_1\,s + p_0$$

is prescribed, one must choose

$$\begin{aligned} \boldsymbol{\ell}^{\mathrm{T}} &= p_0\,\mathbf{q}_C^{\mathrm{T}} + p_1\,\mathbf{q}_C^{\mathrm{T}}\mathbf{A} + \cdots + p_{n-1}\,\mathbf{q}_C^{\mathrm{T}}\,\mathbf{A}^{n-1} + \mathbf{q}_C^{\mathrm{T}}\,\mathbf{A}^n \\ &= \mathbf{q}_C^{\mathrm{T}}\,p(\mathbf{A}) \end{aligned}\ , \tag{1.27}$$

where $\mathbf{q}_C^{\mathrm{T}}$ *is the last row of the controllability matrix inverse,* $\mathbf{Q}_C^{-1}$*, and is thus* *determined, according to (A.33) of Appendix A, by*

$$\mathbf{q}_C^{\mathrm{T}} = \begin{pmatrix} 0 & \cdots & 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{b} & \mathbf{A}\mathbf{b} & \cdots & \mathbf{A}^{n-1}\mathbf{b} \end{pmatrix}^{-1}.$$

This is the *Ackermann's theorem*, and (1.27) is the *Ackermann's formula*.

1.2.2.3 Example

Consider the single input, single output system of third order of Fig. 1.4. Let us design the state-feedback control of this system, i.e. calculate the constants ℓ_1, ℓ_2 and ℓ_3, which are the elements of its (1×3) feedback matrix $\boldsymbol{\ell}^{\mathrm{T}} = (\ell_1 \quad \ell_2 \quad \ell_3)$, and determine thus its control law $u = -\boldsymbol{\ell}^{\mathrm{T}}\mathbf{x}$. The following poles (eigenvalues) should be imposed to the closed-loop: $s_{1,2} = -2 \pm j2$, $s_3 = -10$. Both methods described above will be used.

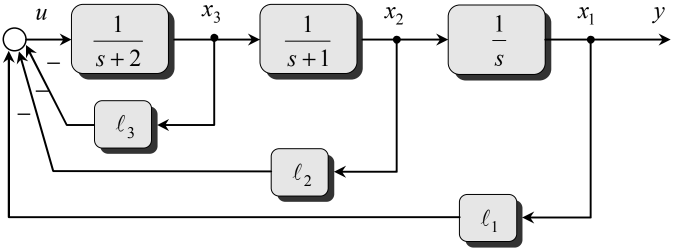


Fig. 1.4 State-feedback control of a third order system.

Direct Calculation: The system order being here sufficiently small $(n = 3)$, the system of equations (1.19) can be solved directly. The state-variable representation of this system by means of the state variables indicated in the figure is obtained by direct inspection:

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}u \\ y = \mathbf{c}^{\mathrm{T}}\mathbf{x} \end{cases}$$

with

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -2 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{c}^{\mathrm{T}} = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}.$$

Let us calculate the characteristic polynomial of the closed-loop system. Since

$$\mathbf{b}\boldsymbol{\ell}^{\mathrm{T}}=\begin{pmatrix}0\\0\\1\end{pmatrix}\begin{pmatrix}\ell_1 & \ell_2 & \ell_3\end{pmatrix}=\begin{pmatrix}0 & 0 & 0\\0 & 0 & 0\\ \ell_1 & \ell_2 & \ell_3\end{pmatrix},$$

the closed-loop system matrix is given by

$$\mathbf{A}-\mathbf{b}\boldsymbol{\ell}^{\mathrm{T}}=\begin{pmatrix}0 & 1 & 0\\0 & -1 & 1\\ -\ell_1 & -\ell_2 & -2-\ell_3\end{pmatrix},$$

and the corresponding characteristic polynomial has the following expression:

$$\det\left[s\mathbf{I}-(\mathbf{A}-\mathbf{b}\boldsymbol{\ell}^{\mathrm{T}})\right]=\begin{vmatrix}s & -1 & 0\\0 & s+1 & -1\\ \ell_1 & \ell_2 & s+2+\ell_3\end{vmatrix}=s^3+(\ell_3+3)s^2+(\ell_2+\ell_3+2)s+\ell_1 .$$

By coefficient identification of this polynomial with those of the prescribed one, namely

$$\begin{aligned}p(s)&=(s+2+j2)(s+2-j2)(s+10)=(s^2+4s+8)(s+10)\\&=s^3+14s^2+48s+80=s^3+p_2s^2+p_1s+p_0\ ,\end{aligned}$$

one obtains readily:

$$\begin{aligned}p_2&=14=\ell_3+3\,,\\p_1&=48=\ell_2+\ell_3+2\,,\\p_0&=80=\ell_1\,,\end{aligned}$$

which yields, by solving this system of linear equations in the unknowns ℓ_1, ℓ_2 and ℓ_3:

$$\ell_1=80\,,\quad \ell_2=48-11-2=35\,,\quad \ell_3=14-3=11\,.$$

Finally:

$$u=-80x_1-35x_2-11x_3\,.$$

Use of Ackermann’s formula: Calculate successively

$$\mathbf{Ab} = \begin{pmatrix}0 & 1 & 0\\ 0 & -1 & 1\\ 0 & 0 & -2\end{pmatrix}\begin{pmatrix}0\\ 0\\ 1\end{pmatrix} = \begin{pmatrix}0\\ 1\\ -2\end{pmatrix} \text{ and } \mathbf{A}^2\mathbf{b} = \mathbf{A}\cdot\mathbf{Ab} = \begin{pmatrix}0 & 1 & 0\\ 0 & -1 & 1\\ 0 & 0 & -2\end{pmatrix}\begin{pmatrix}0\\ 1\\ -2\end{pmatrix} = \begin{pmatrix}1\\ -3\\ 4\end{pmatrix},$$

and determine then the row vector $\mathbf{q}_C^{\mathrm{T}} = (q_1 \quad q_2 \quad q_3)$ according to (A.33):

$$\begin{aligned}
&\mathbf{q}_C^{\mathrm{T}}\mathbf{b} = 0 && \Rightarrow \quad q_3 = 0\,,\\
&\mathbf{q}_C^{\mathrm{T}}\mathbf{Ab} = 0 && \Rightarrow \quad q_2 - 2q_3 = 0 \quad \Rightarrow \quad q_2 = 0\,,\\
&\mathbf{q}_C^{\mathrm{T}}\mathbf{A}^2\mathbf{b} = 1 && \Rightarrow \quad q_1 - 3q_2 + 4q_3 = 1 \quad \Rightarrow \quad q_1 = 1\,.
\end{aligned}$$

Therefore:

$$\begin{array}{l|l}
\mathbf{q}_C^{\mathrm{T}} = \begin{pmatrix}1 & 0 & 0\end{pmatrix} & \times p_0 = 80\times\begin{pmatrix}1 & 0 & 0\end{pmatrix}\\
\mathbf{q}_C^{\mathrm{T}}\mathbf{A} = \begin{pmatrix}1 & 0 & 0\end{pmatrix}\begin{pmatrix}0 & 1 & 0\\ 0 & -1 & 1\\ 0 & 0 & -2\end{pmatrix} = \begin{pmatrix}0 & 1 & 0\end{pmatrix} & \times p_1 = 48\times\begin{pmatrix}0 & 1 & 0\end{pmatrix}\\
\mathbf{q}_C^{\mathrm{T}}\mathbf{A}^2 = \mathbf{q}_C^{\mathrm{T}}\mathbf{A}\cdot\mathbf{A} = \begin{pmatrix}0 & 1 & 0\end{pmatrix}\begin{pmatrix}0 & 1 & 0\\ 0 & -1 & 1\\ 0 & 0 & -2\end{pmatrix} = \begin{pmatrix}0 & -1 & 1\end{pmatrix} & \times p_2 = 14\times\begin{pmatrix}0 & -1 & 1\end{pmatrix}\\
\mathbf{q}_C^{\mathrm{T}}\mathbf{A}^3 = \mathbf{q}_C^{\mathrm{T}}\mathbf{A}^2\cdot\mathbf{A} = \begin{pmatrix}0 & -1 & 1\end{pmatrix}\begin{pmatrix}0 & 1 & 0\\ 0 & -1 & 1\\ 0 & 0 & -2\end{pmatrix} = \begin{pmatrix}0 & 1 & -3\end{pmatrix} & \times 1 = \qquad \begin{pmatrix}0 & 1 & -3\end{pmatrix}
\end{array}$$

From (1.27),

$$\begin{aligned}
\boldsymbol{\ell}^{\mathrm{T}} &= p_0\mathbf{q}_C^{\mathrm{T}} + p_1\mathbf{q}_C^{\mathrm{T}}\mathbf{A} + p_2\mathbf{q}_C^{\mathrm{T}}\mathbf{A}^2 + \mathbf{q}_C^{\mathrm{T}}\mathbf{A}^3\\
&= \begin{pmatrix}80 & 48-14+1 & 14-3\end{pmatrix} = \begin{pmatrix}80 & 35 & 11\end{pmatrix}.
\end{aligned}$$

This result is identical to the previous one.

1.2.2.4 Case of the Discrete-time Systems

The pole placement method described above applies also to discrete-time systems. As a matter of fact, it is for this type of systems that it has been established initially [Ack72]. We will therefore here only state this theorem. The choice of the closed-loop eigenvalues can result from a simple transposition of the values λ_{Li}

chosen for a continuous-time design, by the relation $\nu_{Li} = e^{\lambda_{Li} T_s}$, where T_s is the period at which the continuous system has been sampled ([AsWi97], [GoOs03]).

Theorem 1.3. *If the single-input plant* $\mathbf{x}_{k+1} = \mathbf{\Phi} \mathbf{x}_k + \boldsymbol{\gamma} u_k$ *is controllable, and if it is desired that the state-feedback control law* $u_k = -\boldsymbol{\ell}^{\mathrm{T}} \mathbf{x}_k$ *yields the closed-loop system characteristic polynomial*

$$p(z) = z^n + p_{n-1} z^{n-1} + \cdots + p_1 z + p_0 = (z - \nu_{L1}) \cdots (z - \nu_{Ln}),$$

one must choose

$$\begin{aligned} \boldsymbol{\ell}^{\mathrm{T}} &= p_0 \mathbf{q}_C^{\mathrm{T}} + p_1 \mathbf{q}_C^{\mathrm{T}} \mathbf{\Phi} + \cdots + p_{n-1} \mathbf{q}_C^{\mathrm{T}} \mathbf{\Phi}^{n-1} + \mathbf{q}_C^{\mathrm{T}} \mathbf{\Phi}^n \\ &= \mathbf{q}_C^{\mathrm{T}} p(\mathbf{\Phi}) \end{aligned} , \tag{1.28}$$

where $\mathbf{q}_C^{\mathrm{T}}$ *is the last row of the controllability matrix inverse,* $\mathbf{Q}_C^{-1}$*, and is thus determined by:* $\mathbf{q}_C^{\mathrm{T}} = \begin{pmatrix} 0 & \cdots & 0 & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{\gamma} & \mathbf{\Phi}\boldsymbol{\gamma} & \cdots & \mathbf{\Phi}^{n-1}\boldsymbol{\gamma} \end{pmatrix}^{-1}$.

1.2.3 Multiple-input Systems

The problem of determining **L** remains solvable if the plant is controllable, but, as seen in Sect. 1.2.1, the solution is no longer unique. This can be put to advantage to impose additional conditions to the closed-loop system.

In the methods which will be presented in Sections 1.4 through 1.7, in addition to the wished eigenvalues a particular *structure* will be imposed to the closed loop.

1.2.4 Plant Zeros and Closed-loop Zeros

The previous theorems for SISO systems, as well as those which will be presented later concerning MIMO systems, allow shifting the poles (actually, the eigenvalues) of a system to any point of the complex plane by means of a constant, suitably chosen, state feedback, if this system is controllable.

The question which arises then is the following: what happens to the plant zeros during this pole-shifting process? That question is of utmost importance, since, if the zeros move arbitrarily, they can deteriorate the dynamic behavior, which was aimed at by choosing a new pole configuration. An extreme case, e.g., would be that of a zero falling accidentally at the same place as one of the shifted poles.

Theorem 1.4. *The open-loop zeros are not modified by the pole placement procedure. In other words, the zeros of a system, which is closed by a state-feedback matrix* **L**, *are the zeros of the initial plant.*

Proof. The zeros of a continuous-time system, for which it is assumed that $p = q$ and $\mathbf{D} = 0$, and which is closed by a state feedback, are the s-values solving the following equation, see Appendix A, (A.29):

$$\begin{vmatrix} s\mathbf{I} - \mathbf{A} + \mathbf{BL} & -\mathbf{B} \\ \mathbf{C} & 0 \end{vmatrix} = 0 .$$

The value of this $(n+p)\times(n+q)$ determinant remains unchanged if one adds to the n first columns the last column, right multiplied by the **L** matrix:

$$\begin{vmatrix} s\mathbf{I} - \mathbf{A} & -\mathbf{B} \\ \mathbf{C} & 0 \end{vmatrix} = 0 .$$

According to (A.29) again, this equation yields the zeros of the initial plant. There is thus identity between these zeros and those of the closed-loop system.

1.3 Finite Settling-time Design for Discrete Systems

1.3.1 Problem Description

This design method, also called dead-beat control, applies only to discrete-time systems. Consider such a system with a scalar input,

$$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \boldsymbol{\gamma} u_k ,$$

where $\mathbf{\Phi}$ is of size $(n \times n)$.

The goal is to find a control sequence, or control law, $\{u_o, u_1, \cdots, u_k\}$, which transfers the system from an arbitrary initial state $\mathbf{x}_0$ into the final state $\mathbf{x}_f = 0$ (the origin) in the minimum number of discrete time steps.

1.3.2 Solution: Algorithm of Mullin and De Barbeyrac

This problem can be solved in n discrete time steps at most, provided the system is controllable and the state-feedback control is possible [MuBa64].

Proof. Let $u_k = -\boldsymbol{\ell}^{\mathrm{T}}\mathbf{x}_k$ be the searched control law. In order to establish it, let us calculate the state at successive discrete times:

$\mathbf{x}_0$ (arbitrary initial state)

$$\mathbf{x}_1 = \boldsymbol{\Phi}\mathbf{x}_0 + \boldsymbol{\gamma}u_0 = \boldsymbol{\Phi}\mathbf{x}_0 - \boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}}\mathbf{x}_0 = (\boldsymbol{\Phi} - \boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}})\mathbf{x}_0$$

$$\begin{aligned}
\mathbf{x}_2 = \boldsymbol{\Phi}\mathbf{x}_1 + \boldsymbol{\gamma}u_1 &= \underbrace{\boldsymbol{\Phi}\mathbf{x}_1 - \boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}}\mathbf{x}_1}_{} = (\boldsymbol{\Phi} - \boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}})\mathbf{x}_1 = (\boldsymbol{\Phi} - \boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}})^2\,\mathbf{x}_0 \\
&= \boldsymbol{\Phi}\,(\boldsymbol{\Phi} - \boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}})\mathbf{x}_0 - \boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}}(\boldsymbol{\Phi} - \boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}})\mathbf{x}_0 \\
&= \left\{\boldsymbol{\Phi}^2 - \begin{pmatrix}\boldsymbol{\gamma} & \boldsymbol{\Phi}\boldsymbol{\gamma}\end{pmatrix}\begin{pmatrix}\boldsymbol{\ell}^{\mathrm{T}}(\boldsymbol{\Phi} - \boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}}) \\ \boldsymbol{\ell}^{\mathrm{T}}\end{pmatrix}\right\}\mathbf{x}_0
\end{aligned}$$

$$\begin{aligned}
\mathbf{x}_3 = \boldsymbol{\Phi}\mathbf{x}_2 + \boldsymbol{\gamma}u_2 &= \underbrace{\boldsymbol{\Phi}\mathbf{x}_2 - \boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}}\mathbf{x}_2}_{} = (\boldsymbol{\Phi} - \boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}})\mathbf{x}_2 = (\boldsymbol{\Phi} - \boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}})^3\,\mathbf{x}_0 \\
&= \left\{\boldsymbol{\Phi}\left\{\boldsymbol{\Phi}^2 - \begin{pmatrix}\boldsymbol{\gamma} & \boldsymbol{\Phi}\boldsymbol{\gamma}\end{pmatrix}\begin{pmatrix}\boldsymbol{\ell}^{\mathrm{T}}(\boldsymbol{\Phi} - \boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}}) \\ \boldsymbol{\ell}^{\mathrm{T}}\end{pmatrix}\right\} - \boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}}(\boldsymbol{\Phi} - \boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}})^2\right\}\mathbf{x}_0 \\
&= \left\{\boldsymbol{\Phi}^3 - \begin{pmatrix}\boldsymbol{\Phi}\boldsymbol{\gamma} & \boldsymbol{\Phi}^2\boldsymbol{\gamma}\end{pmatrix}\begin{pmatrix}\boldsymbol{\ell}^{\mathrm{T}}(\boldsymbol{\Phi} - \boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}}) \\ \boldsymbol{\ell}^{\mathrm{T}}\end{pmatrix} - \boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}}(\boldsymbol{\Phi} - \boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}})^2\right\}\mathbf{x}_0 \\
&= \left\{\boldsymbol{\Phi}^3 - \begin{pmatrix}\boldsymbol{\gamma} & \boldsymbol{\Phi}\boldsymbol{\gamma} & \boldsymbol{\Phi}^2\boldsymbol{\gamma}\end{pmatrix}\begin{pmatrix}\boldsymbol{\ell}^{\mathrm{T}}(\boldsymbol{\Phi} - \boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}})^2 \\ \boldsymbol{\ell}^{\mathrm{T}}(\boldsymbol{\Phi} - \boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}}) \\ \boldsymbol{\ell}^{\mathrm{T}}\end{pmatrix}\right\}\mathbf{x}_0,\ \dots
\end{aligned}$$

and so on. We can extrapolate easily this result to step n:

$$\mathbf{x}_n = \left\{\boldsymbol{\Phi}^n - \begin{pmatrix}\boldsymbol{\gamma} & \boldsymbol{\Phi}\boldsymbol{\gamma} & \cdots & \boldsymbol{\Phi}^{n-1}\boldsymbol{\gamma}\end{pmatrix}\begin{pmatrix}\boldsymbol{\ell}^{\mathrm{T}}(\boldsymbol{\Phi} - \boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}})^{n-1} \\ \vdots \\ \boldsymbol{\ell}^{\mathrm{T}}(\boldsymbol{\Phi} - \boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}}) \\ \boldsymbol{\ell}^{\mathrm{T}}\end{pmatrix}\right\}\mathbf{x}_0\,.$$

The state at this step will be equal to the state space origin (the final state), $\mathbf{x}_n = 0$, and this $\forall\,\mathbf{x}_0$, if and only if

$$\left\{\boldsymbol{\Phi}^n - \begin{pmatrix}\boldsymbol{\gamma} & \boldsymbol{\Phi\gamma} & \cdots & \boldsymbol{\Phi}^{n-1}\boldsymbol{\gamma}\end{pmatrix}\begin{pmatrix}\boldsymbol{\ell}^{\mathrm{T}}(\boldsymbol{\Phi}-\boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}})^{n-1}\\ \vdots \\ \boldsymbol{\ell}^{\mathrm{T}}(\boldsymbol{\Phi}-\boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}})\\ \boldsymbol{\ell}^{\mathrm{T}}\end{pmatrix}\right\} = 0\,,$$

or:

$$\begin{pmatrix}\boldsymbol{\ell}^{\mathrm{T}}(\boldsymbol{\Phi}-\boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}})^{n-1}\\ \vdots \\ \boldsymbol{\ell}^{\mathrm{T}}(\boldsymbol{\Phi}-\boldsymbol{\gamma}\boldsymbol{\ell}^{\mathrm{T}})\\ \boldsymbol{\ell}^{\mathrm{T}}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\gamma} & \boldsymbol{\Phi\gamma} & \cdots & \boldsymbol{\Phi}^{n-1}\boldsymbol{\gamma}\end{pmatrix}^{-1}\boldsymbol{\Phi}^n\,.$$

The matrix to be inverted in the second member of this equation is nothing else but the controllability matrix $\mathbf{Q}_C$ of the system (see (A.13), Appendix A). Its inverse exists therefore if the system is controllable, hypothesis which has been introduced at the beginning of this section. The final result is thus the following: The desired control law $\boldsymbol{\ell}^{\mathrm{T}}$ is the last row of the matrix

$$\begin{pmatrix}\boldsymbol{\gamma} & \boldsymbol{\Phi\gamma} & \cdots & \boldsymbol{\Phi}^{n-1}\boldsymbol{\gamma}\end{pmatrix}^{-1}\boldsymbol{\Phi}^n\,,$$

i.e., with $\mathbf{q}_C^{\mathrm{T}}$ being defined as the last row of $\mathbf{Q}_C^{-1}$ (see Sect. A.7.2):

$$\boldsymbol{\ell}^{\mathrm{T}} = \mathbf{q}_C^{\mathrm{T}}\,\boldsymbol{\Phi}^n \tag{1.29}$$

Remark 1.5. In the general case, a sequence of n steps is thus necessary for this control law to bring the system to its final state. However, depending on the initial state, it may happen that less than n steps are sufficient.

Remark 1.6. The remarkable result obtained derives also from Ackermann's theorem, as a special case, by choosing to place all the discrete system poles at the origin of the z-plane ($\nu_i = 0, \quad \forall\, i = 1,\ldots,n$), i.e. by prescribing the characteristic polynomial

$$p(z) = z^n\,.$$

What we have obtained here by a different derivation is the same as the special *finite settling time* variant of a general RST structure [GoOs03].

Remark 1.7. This type of control law, due to the fact that it drives the control system to equilibrium in a minimum number of discrete time steps, provides frequently control signals u_k of high level, at least during the first steps. For this reason it has acquired, according to [AsWi97], "*an undeservedly bad reputation*". In the case of the digital control of a real continuous-time plant, it is indeed often possible to apply this method with some sacrifice, e.g. by increasing the sampling time T_s to such a value that the first control signal does no longer exceed the allowed limit, if the choice of T_s is not dictated by other considerations. Another possibility, if the sampling time cannot be adjusted, consists in imposing a constraint to this first control signal, which obliges then to give up the control in a *minimum number of steps*, and to add one or several sampling steps beyond n.

1.4 Simple (or Reduced) Modal Control

1.4.1 Definition

By the similitude transformation

$$\mathbf{x} = \mathbf{T}\mathbf{x}^* = x_1^*\,\mathbf{v}_1 + x_2^*\,\mathbf{v}_2 + \cdots + x_n^*\,\mathbf{v}_n\,,$$

where $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n$ are the eigenvectors of $\mathbf{A}$, it is possible to diagonalize the system matrix $\mathbf{A}$ if all its eigenvalues are distinct (see Sect. A.7.4, Appendix A).

In the new basis, the basis of the eigenvectors, the state differential equations

$$\dot{x}_i^* = \lambda_i\, x_i^* + u_i^*$$

are then entirely decoupled. If the plant has $p = n$ inputs and if the control inputs u_i^* are independent from each other, each state variable x_i^*, and consequently each *mode* $x_i^*\,\mathbf{v}_i$, can be controlled separately. This is why this control is referred to as *modal control.*

In practice, it will generally only be possible to come close to an ideal modal control, first because there may exist multiple eigenvalues and the state variables associated with them are not completely decoupled, second because usually the plant does not have as many control inputs as state variables.

1.4.2 Simple Modal Control Design Method

In the remainder of this section, the eigenvalues will be assumed all distinct. Since p control inputs are available, we will try to control individually p modal coordinates. Due to the formal similarity of the expressions, only the continuous case will be considered here.

The plant state representation, in its most general form

$$\dot{\mathbf{x}} = \mathbf{A}\,\mathbf{x} + \mathbf{B}\,\mathbf{u}\,,$$

where $\text{size}(\mathbf{A}) = n \times n$ and $\text{size}(\mathbf{u}) = p \times 1$ with $p \leq n$, is transformed into its diagonal form

$$\dot{\mathbf{x}}^* = \mathbf{\Lambda}\,\mathbf{x}^* + \hat{\mathbf{B}}\,\mathbf{u} \tag{1.30}$$

by the transformation

$$\mathbf{x} = \mathbf{T}\,\mathbf{x}^*\,,$$

with

$$\mathbf{\Lambda} = \mathbf{T}^{-1}\mathbf{A}\,\mathbf{T}$$

and

$$\hat{\mathbf{B}} = \mathbf{T}^{-1}\,\mathbf{B} = \begin{pmatrix} \mathbf{w}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{w}_n^{\mathrm{T}} \end{pmatrix} \mathbf{B} = \begin{pmatrix} \mathbf{w}_1^{\mathrm{T}}\,\mathbf{B} \\ \vdots \\ \mathbf{w}_n^{\mathrm{T}}\,\mathbf{B} \end{pmatrix}, \tag{1.31}$$

where $\mathbf{T}$ is the modal matrix and where $\mathbf{w}_i^{\mathrm{T}}$ denotes the i-th row of its inverse $\mathbf{T}^{-1}$ (see (A.30) and Sect. A.7.4 of Appendix A).

1.4.2.1 Choice of the *p* Modal Coordinates Fed Back to the Input

This choice depends on the real data of the plant to control. It is common to use for the feedback control those coordinates which have associated eigenvalues closest to the imaginary axis, since they are the ones which determine principally the dynamic behavior of the plant and must therefore be controlled in priority.

In cases where essentially the plant behavior with respect to a disturbance is of interest, the modal coordinates which are most influenced by this disturbance should be retained.

Let us group them together in a vector $\mathbf{x}_p^* = \begin{pmatrix} x_1^* & \cdots & x_p^* \end{pmatrix}^{\mathrm{T}}$, and call $\mathbf{x}_{n-p}^*$ the vector of remaining modal coordinates. (1.30) can then be written as

$$\begin{pmatrix} \dot{\mathbf{x}}_p^* \\ \dot{\mathbf{x}}_{n-p}^* \end{pmatrix} = \begin{pmatrix} \mathbf{\Lambda}_p & 0 \\ 0 & \mathbf{\Lambda}_{n-p} \end{pmatrix} \begin{pmatrix} \mathbf{x}_p^* \\ \mathbf{x}_{n-p}^* \end{pmatrix} + \begin{pmatrix} \hat{\mathbf{B}}_p \\ \hat{\mathbf{B}}_{n-p} \end{pmatrix} \mathbf{u}\,, \tag{1.32}$$

where: $\mathbf{\Lambda}_p$ $= (p \times p)$ north-west submatrix of $\mathbf{\Lambda}$;
$\mathbf{\Lambda}_{n-p}$ $= (n-p) \times (n-p)$ south-east submatrix of $\mathbf{\Lambda}$;
$\hat{\mathbf{B}}_p$ $=$ submatrix of $\hat{\mathbf{B}}$ made of its first p rows, thus of size $(p \times p)$;
$\hat{\mathbf{B}}_{n-p}$ $=$ submatrix of $\hat{\mathbf{B}}$ made of its last $(n-p)$ rows.

This yields

$$\begin{cases} \dot{\mathbf{x}}_p^* &= \mathbf{\Lambda}_p\, \mathbf{x}_p^* + \hat{\mathbf{B}}_p\, \mathbf{u} \\ \dot{\mathbf{x}}_{n-p}^* &= \mathbf{\Lambda}_{n-p}\, \mathbf{x}_{n-p}^* + \hat{\mathbf{B}}_{n-p}\, \mathbf{u} \end{cases} \tag{1.33}$$

The first of these two equations can be written

$$\dot{\mathbf{x}}_p^* = \mathbf{\Lambda}_p\, \mathbf{x}_p^* + \mathbf{u}^*\,, \tag{1.34}$$

by introducing

$$\mathbf{u}^* = \hat{\mathbf{B}}_p\, \mathbf{u}\,. \tag{1.35}$$

1.4.2.2 Principle of the Simple Modal Control

We want now to determine the feedback of these p first modal coordinates in such a way that, simultaneously,

1. the given eigenvalues $\lambda_1, \cdots, \lambda_p$ are replaced by the desired eigenvalues: $\lambda_{L1}, \cdots, \lambda_{Lp}$;
2. the closed-loop system remains diagonal as far as these p state variables are concerned; this is the principle of modal control: each mode is controlled independently.

The state equation of the closed-loop system, in regulation behavior $(\mathbf{y}_r = 0)$, must therefore have the following form for the p involved coordinates:

$$\dot{\mathbf{x}}_p^* = \mathbf{\Lambda}_{Lp}\, \mathbf{x}_p^*\,, \text{ where } \mathbf{\Lambda}_{Lp} = \begin{pmatrix} \lambda_{L1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_{Lp} \end{pmatrix},$$

and, consequently, the following equality must hold:

$$\mathbf{\Lambda}_p \, \mathbf{x}_p^* + \mathbf{u}^* = \mathbf{\Lambda}_{Lp} \, \mathbf{x}_p^* ,$$

yielding

$$\mathbf{u}^* = -(\mathbf{\Lambda}_p - \mathbf{\Lambda}_{Lp}) \, \mathbf{x}_p^* = -\begin{pmatrix} \lambda_1 - \lambda_{L1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p - \lambda_{Lp} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1^* \\ \vdots \\ \mathbf{x}_p^* \end{pmatrix} . \tag{1.36}$$

Note that $\mathbf{u}^*$ is a fictitious control vector, which does not exist in the real system. It depends on **u** according to (1.35). This equation must be solvable in **u**, which is the real control vector acting on the plant:

$$\mathbf{u} = \hat{\mathbf{B}}_p^{-1} \, \mathbf{u}^* . \tag{1.37}$$

For that purpose, $\hat{\mathbf{B}}_p$, which is a square $(p \times p)$ matrix, must be regular. Now $\hat{\mathbf{B}}_p$ is a submatrix made from the *p* first rows of $\hat{\mathbf{B}}$, which is itself an $(n \times p)$ matrix of rank *p*, since **B** has this rank and since $\mathbf{T}^{-1}$ is a regular matrix (thus composed of linearly independent rows $\mathbf{w}_i^{\mathrm{T}}$). It is therefore possible to find in $\widehat{\mathbf{B}}$ a number *p* of linearly independent rows. If these rows are not the first ones, it is always possible to renumber the state variables for this to be the case. Of course, if a set of *p* modal coordinates to be controlled has already been chosen, it may then be necessary to modify this choice and, eventually, to drop one or another of these variables for feedback control.

Equations (1.36) and (1.37) yield

$$\mathbf{u} = -\hat{\mathbf{B}}_p^{-1} \, (\mathbf{\Lambda}_p - \mathbf{\Lambda}_{Lp}) \, \mathbf{x}_p^* , \tag{1.38}$$

then, by substitution into the differential equations (1.33),

$$\begin{cases} \dot{\mathbf{x}}_p^* &= \mathbf{\Lambda}_{Lp} \, \mathbf{x}_p^* \\ \dot{\mathbf{x}}_{n-p}^* &= -\hat{\mathbf{B}}_{n-p} \hat{\mathbf{B}}_p^{-1} \, (\mathbf{\Lambda}_p - \mathbf{\Lambda}_{Lp}) \, \mathbf{x}_p^* + \mathbf{\Lambda}_{n-p} \, \mathbf{x}_{n-p}^* \end{cases} \tag{1.39}$$

The first equation (1.39) shows that the *p* first modal coordinates are indeed controlled individually. The corresponding eigenvalues, λ_{L1} to λ_{Lp}, can be imposed arbitrarily.

1.4.2.3 Influence on the Other Modal Coordinates

The second equation (1.39) shows that the $(n-p)$ remaining modal coordinates are influenced by the p first ones. It is therefore no longer possible to control them arbitrarily.

What can we say then about the closed-loop stability? The closed-loop eigenvalues are obtained by solving

$$\begin{vmatrix} s\,\mathbf{I}_p - \mathbf{\Lambda}_{Lp} & 0 \\ \hat{\mathbf{B}}_{n-p}\,\hat{\mathbf{B}}_p^{-1}(\mathbf{\Lambda}_p - \mathbf{\Lambda}_{Lp}) & s\,\mathbf{I}_{n-p} - \mathbf{\Lambda}_{n-p} \end{vmatrix} = 0\,,$$

thus

$$\left| s\,\mathbf{I}_p - \mathbf{\Lambda}_{Lp} \right| \times \left| s\,\mathbf{I}_{n-p} - \mathbf{\Lambda}_{n-p} \right| = 0\,,$$

i.e., explicitly

$$(s-\lambda_{L1})\cdots(s-\lambda_{Lp})(s-\lambda_{p+1})\cdots(s-\lambda_n) = 0\,.$$

This means that the remaining $(n-p)$ closed-loop eigenvalues are equal to the plant eigenvalues $\lambda_{p+1},\cdots,\lambda_n$, thus have remained unchanged, whereas the p first ones have been shifted from λ_i to λ_{Li} by the simple modal feedback.

Conclusion: the closed-loop stability is not endangered by the repercussion of the control of the p first modal coordinates on the remaining $(n-p)$ ones.

1.4.2.4 Simple Modal Controller

The entire previous development has taken place in the state space spanned by the eigenvectors. We must now go back to the initial state representation. With

$$\mathbf{x}^* = \mathbf{T}^{-1}\,\mathbf{x} = \begin{pmatrix} \mathbf{w}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{w}_n^{\mathrm{T}} \end{pmatrix} \mathbf{x}\,,$$

we get from (1.38):

$$\mathbf{u} = -\hat{\mathbf{B}}_p^{-1}\,(\mathbf{\Lambda}_p - \mathbf{\Lambda}_{Lp}) \begin{pmatrix} \mathbf{w}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{w}_p^{\mathrm{T}} \end{pmatrix} \mathbf{x} = -\mathbf{L}\,\mathbf{x}\,.$$

The simple modal controller is thus given by the $(p \times n)$ matrix

$$\mathbf{L} = \begin{pmatrix} \mathbf{w}_1^{\mathrm{T}}\,\mathbf{B} \\ \vdots \\ \mathbf{w}_p^{\mathrm{T}}\,\mathbf{B} \end{pmatrix}^{-1} \begin{pmatrix} \lambda_1 - \lambda_{L1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p - \lambda_{Lp} \end{pmatrix} \begin{pmatrix} \mathbf{w}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{w}_p^{\mathrm{T}} \end{pmatrix}, \tag{1.40}$$

where the employed notations have the following recalled significations:

$\mathbf{w}_i^{\mathrm{T}}$ = i-th row of the inverse transformation matrix, $\mathbf{T}^{-1}$,

$\left\{\lambda_i\right\}_{i=1,\ldots,p}$ = p first plant eigenvalues,

$\left\{\lambda_{Li}\right\}_{\ddot{\imath}=1,\ldots,p}$ = p first closed-loop eigenvalues (prescribed values).

Remark 1.8. The present control is again a proportional control law.

1.4.2.5 Block Diagram

The previous modal control applied to the state vector is represented in Fig. 1.5.

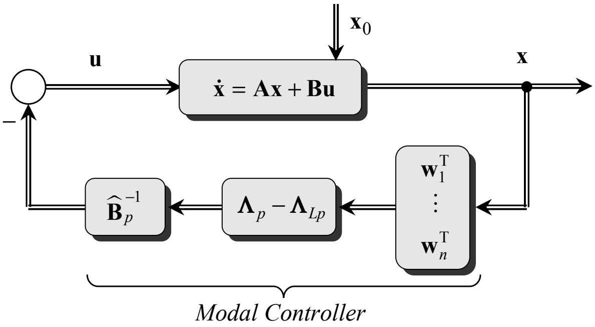

$\mathbf{x}_0$
$\mathbf{u}$
$\mathbf{x}$
$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$
$-$
$\widehat{\mathbf{B}}_p^{-1}$
$\mathbf{\Lambda}_p - \mathbf{\Lambda}_{Lp}$
$\mathbf{w}_1^{\mathrm{T}}$
$\vdots$
$\mathbf{w}_n^{\mathrm{T}}$
Modal Controller

Fig. 1.5 State feedback by modal control.

The drawback of this diagram is that it does not show the decoupling of the p first modal coordinates. Let us thus transpose the previous equations to the s-plane.

1.4.2.6 Application of the Laplace Transform

Applying the Laplace transform to the two equations (1.39),

$$\begin{cases} s\mathbf{X}_p^*(s) - \mathbf{x}_p^*(0) = \mathbf{\Lambda}_{Lp}\ \mathbf{X}_p^*(s) \\ s\mathbf{X}_{n-p}^*(s) - \mathbf{x}_{n-p}^*(0) = -\hat{\mathbf{B}}_{n-p}\ \hat{\mathbf{B}}_p^{-1}(\mathbf{\Lambda}_p - \mathbf{\Lambda}_{Lp})\mathbf{X}_p^*(s) + \mathbf{\Lambda}_{n-p}\ \mathbf{X}_{n-p}^*(s) \end{cases}$$

where $\mathbf{X}_p^*(s)=\mathcal{L}\left[\mathbf{x}_p^*(t)\right]$ and $\mathbf{X}_{n-p}^*(s)=\mathcal{L}\left[\mathbf{x}_{n-p}^*(t)\right]$, we obtain:

$$\begin{cases}\mathbf{X}_p^*(s)=(s\mathbf{I}-\mathbf{\Lambda}_{Lp})^{-1}\,\mathbf{x}_p^*(0)\\ \mathbf{X}_{n-p}^*(s)=(s\mathbf{I}-\mathbf{\Lambda}_{n-p})^{-1}\left[\mathbf{x}_{n-p}^*(0)-\hat{\mathbf{B}}_{n-p}\,\hat{\mathbf{B}}_p^{-1}(\mathbf{\Lambda}_p-\mathbf{\Lambda}_{Lp})\mathbf{X}_p^*(s)\right]\end{cases}$$

Since the matrices to be inverted are diagonal, so are their inverses. They have the diagonal elements

$$\frac{1}{s-\lambda_{Li}}\quad \text{for } i=1,\ \dots\ ,p \quad \text{and} \quad \frac{1}{s-\lambda_i}\quad \text{for } i=p+1,\dots,n\,.$$

Taking into account the fact that

$$\mathbf{x}^*(0)=\mathbf{T}^{-1}\,\mathbf{x}(0)=\begin{pmatrix}\mathbf{w}_1^{\mathrm{T}}\\ \vdots\\ \mathbf{w}_n^{\mathrm{T}}\end{pmatrix}\mathbf{x}(0)\,,$$

the block diagram of Fig. 1.6 is obtained.

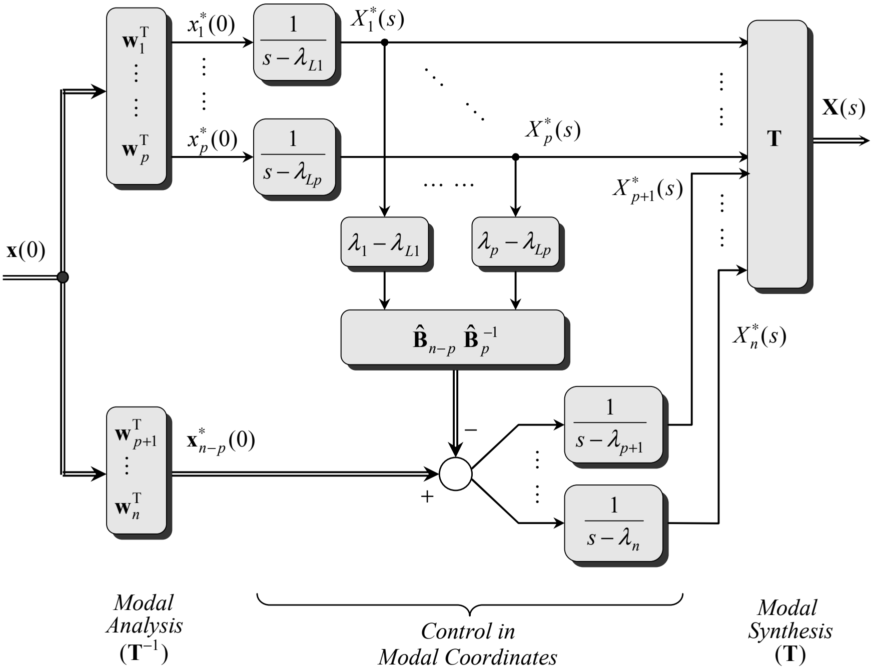

$\mathbf{x}(0)$
$\mathbf{w}_1^{\mathrm{T}}$
$\mathbf{w}_p^{\mathrm{T}}$
$x_1^*(0)$
$x_p^*(0)$
$\frac{1}{s-\lambda_{L1}}$
$\frac{1}{s-\lambda_{Lp}}$
$X_1^*(s)$
$X_p^*(s)$
$X_{p+1}^*(s)$
$X_n^*(s)$
$\lambda_1-\lambda_{L1}$
$\lambda_p-\lambda_{Lp}$
$\hat{\mathbf{B}}_{n-p}\,\hat{\mathbf{B}}_p^{-1}$
$\mathbf{T}$
$\mathbf{X}(s)$
$\mathbf{w}_{p+1}^{\mathrm{T}}$
$\mathbf{w}_n^{\mathrm{T}}$
$\mathbf{x}_{n-p}^*(0)$
−
+
$\frac{1}{s-\lambda_{p+1}}$
$\frac{1}{s-\lambda_n}$
Modal Analysis ($\mathbf{T}^{-1}$)
Control in Modal Coordinates
Modal Synthesis ($\mathbf{T}$)

Fig. 1.6 Diagram of the modal control detailed as to the modal coordinates.

This figure shows clearly that:

- the first p modal coordinates are controlled individually and have acquired thereby the eigenvalues λ_{L1} to λ_{Lp};
- the $(n-p)$ remaining modal coordinates are influenced by the first p ones, but even so, they have not lost their eigenvalues $\lambda_{p+1},\cdots,\lambda_n$.

1.4.2.7 Special Case: *p* = *n*

It is then possible to avoid the modal transformation **T** if one wants only to shift the eigenvalues. By writing, indeed,

$$\dot{\mathbf{x}} = (\mathbf{A}-\mathbf{BL})\,\mathbf{x} = \mathbf{\Lambda}_L\,\mathbf{x}\,,$$

one obtains directly

$$\mathbf{L} = \mathbf{B}^{-1}(\mathbf{A}-\mathbf{\Lambda}_L)\,, \qquad (1.41)$$

since in this case **B** is invertible $(\mathrm{rank}(\mathbf{B}) = p = n)$.

1.5 Input – Output Decoupling Method

1.5.1 Initial Hypotheses

The approach presented here, which has been introduced by [FaWo67], aims at producing a closed-loop system in which each input influences only one output, therefore its denomination of input – output decoupling control. It applies thus only to systems having a number of inputs identical to the number of outputs: $p = q$.

That restriction is perfectly justified: for the case where $p < q$, the decoupling would anyway not be possible, and, for that where $p > q$, it is easy to revert to the case $p = q$ by dropping $(p-q)$ plant inputs for control purposes. It is assumed furthermore that $\mathrm{rank}(\mathbf{B}) = p$ and that $\mathrm{rank}(\mathbf{C}) = q$, i.e. that the plant state model has neither redundant inputs nor redundant outputs, as it should be the case in every state representation (see Equations (A.7) and (A.8) of Appendix A and the associated conditions about the ranks of **B** and **C**). Finally, as will be shown in the sequel, the plant must be not only controllable, but also observable.

1.5.2 Order Differences of a System

The following developments concern a MIMO continuous-time plant, described by a state representation of type (A.7) of Appendix A, with $\mathbf{D}=0$. They are quite similar in the discrete-time case by simple substitution of z to s. The goal is thus to obtain a closed-loop transfer matrix which has the following shape:

$$\mathbf{G}_{CL}(s)=\begin{pmatrix} G_1(s) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & G_q(s) \end{pmatrix}, \tag{1.42}$$

where the q SISO individual transfer functions $G_i(s)$ have the form

$$G_i(s)=\frac{N_i(s)}{D_i(s)}, \tag{1.43}$$

with

$$N_i(s)=k_i=\text{constant}\,, \tag{1.44}$$

and

$$D_i(s)=s^{\delta_i}+q_{i,\delta_i-1}s^{\delta_i-1}+\ldots+q_{i,1}s+q_{i,0}=(s-p_{i,1})\ldots(s-p_{i,\delta_i})\,, \quad (1.45),$$

where $p_{i,j}$ $(j=1,\ldots,\delta_i)$ are the desired poles for $G_i(s)$ and constitute, together with all the other poles for $i=1,\ldots,q$, the set of poles (eigenvalues) imposed to the decoupled closed-loop system which will result from this design.

The restriction (1.44) imposed to the numerator of $G_i(s)$ is easy to understand. It was proved in Sect. 1.2.4, Theorem 1.4, that the plant zeros remain unchanged when its poles (eigenvalues) are shifted by a state feedback. The zeros of $\mathbf{G}_{CL}(s)$ are thus the same as those of the open-loop transfer matrix $\mathbf{G}_{OL}(s)=\mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}$. Therefore, if one of the transfer functions $G_i(s)$ in (1.42) had a zero $s=\eta$, all the elements of the i-th row of $\mathbf{G}_{OL}(s)$ should vanish for that same value of s, which is extremely rare in practice. A second justification of this restriction will be given a posteriori.

The number δ_i of poles contained in $G_i(s)$ is, as to itself, dictated by the following consideration.

Definition 1.1. The *order difference* of the output y_i is defined as the lowest order δ_i of the derivatives of $y_i(t)$, thus of the first one encountered in order of successive derivatives $y_i^{(\delta_i)}(t)$, on which the input vector **u**, thus the reference vector $\mathbf{y}_r$ in the case of a decoupled control, exerts a *direct* action, its action on all the derivatives of lower order occurring only through the state vector **x**.

This value is the smallest integer number δ_i for which holds

$$\left.\begin{aligned} \mathbf{c}_i^{\mathrm{T}}\mathbf{B} &= 0 \\ \mathbf{c}_i^{\mathrm{T}}\mathbf{AB} &= 0 \\ &\vdots \\ \mathbf{c}_i^{\mathrm{T}}\mathbf{A}^{\delta_i-2}\mathbf{B} &= 0 \\ \mathbf{c}_i^{\mathrm{T}}\mathbf{A}^{\delta_i-1}\mathbf{B} &\neq 0 \end{aligned}\right\} \tag{1.46}$$

where $\mathbf{c}_i^{\mathrm{T}}$ is the *i*-th row of the output matrix **C**, thus yielding $y_i = \mathbf{c}_i^{\mathrm{T}}\mathbf{x}$.
The order difference of the MIMO system with state representation **A**, **B**, **C**, is the sum of the order differences of its various outputs:

$$\delta = \delta_1 + \ldots + \delta_q .$$

The proof that (1.46) yields the order difference δ_i of the output $y_i(t)$ consists in differentiating this output successively and applying at each step the corresponding line of (1.46):

$$\begin{aligned} \dot{y}_i &= \mathbf{c}_i^{\mathrm{T}}\dot{\mathbf{x}} = \mathbf{c}_i^{\mathrm{T}}\mathbf{Ax} + \mathbf{c}_i^{\mathrm{T}}\mathbf{Bu} = \mathbf{c}_i^{\mathrm{T}}\mathbf{Ax}, \\ \ddot{y}_i &= \mathbf{c}_i^{\mathrm{T}}\mathbf{A}\dot{\mathbf{x}} = \mathbf{c}_i^{\mathrm{T}}\mathbf{A}^2\mathbf{x} + \mathbf{c}_i^{\mathrm{T}}\mathbf{ABu} = \mathbf{c}_i^{\mathrm{T}}\mathbf{A}^2\mathbf{x}, \\ &\vdots \\ y_i^{(\delta_i-1)} &= \mathbf{c}_i^{\mathrm{T}}\mathbf{A}^{\delta_i-1}\mathbf{x}, \\ y_i^{(\delta_i)} &= \mathbf{c}_i^{\mathrm{T}}\mathbf{A}^{\delta_i}\mathbf{x} + \mathbf{c}_i^{\mathrm{T}}\mathbf{A}^{\delta_i-1}\mathbf{Bu} \end{aligned} \tag{1.47}$$

$y_i^{(\delta_i)}$ is thus well the first derivative on which **u** acts directly. By combining it with those coming from all the other outputs, with their respective order differences for $i = 1,\ldots,q$, in a fictitious output vector $\mathbf{y}^{*\mathrm{T}} = \left(y_1^{(\delta_1)} \;\; \ldots \;\; y_q^{(\delta_q)}\right)^{\mathrm{T}}$, we can write $\mathbf{y}^* = \mathbf{C}^*\mathbf{x} + \mathbf{D}^*\mathbf{u}$, with

$$\mathbf{C}^* = \begin{pmatrix} \mathbf{c}_1^{\mathrm{T}}\mathbf{A}^{\delta_1} \\ \vdots \\ \mathbf{c}_q^{\mathrm{T}}\mathbf{A}^{\delta_q} \end{pmatrix}, \quad \mathbf{D}^* = \begin{pmatrix} \mathbf{c}_1^{\mathrm{T}}\mathbf{A}^{\delta_1-1}\mathbf{B} \\ \vdots \\ \mathbf{c}_q^{\mathrm{T}}\mathbf{A}^{\delta_q-1}\mathbf{B} \end{pmatrix}. \tag{1.48}$$

1.5.3 Decoupled Control According to Falb-Wolovich

The substitution of the general expression (1.2) of a state-feedback plus gain-compensation control, namely $\mathbf{u} = -\mathbf{L}\,\mathbf{x} + \mathbf{M}\,\mathbf{y}_r$, into the last equation of (1.47) yields, for $i = 1,\dots,p$ since, as mentioned above, $p = q$:

$$y_i^{(\delta_i)} = (\mathbf{c}_i^{\mathrm{T}}\mathbf{A}^{\delta_i} - \mathbf{c}_i^{\mathrm{T}}\mathbf{A}^{\delta_i - 1}\mathbf{BL})\mathbf{x} + \mathbf{c}_i^{\mathrm{T}}\mathbf{A}^{\delta_i - 1}\mathbf{BM}\,\mathbf{y}_r\,. \tag{1.49}$$

With the purpose to obtain that the output y_i be influenced only by its attributed reference $y_{r,i}$, it is then logic to let

$$\mathbf{c}_i^{\mathrm{T}}\mathbf{A}^{\delta_i - 1}\mathbf{BM}\,\mathbf{y}_r = k_i\,y_{r,i} = \begin{pmatrix}0 & \cdots & 0 & k_i & 0 & \cdots & 0\end{pmatrix}\mathbf{y}_r\,, \tag{1.50}$$

where k_i is an arbitrary constant, and this for $i = 1,\dots,p$. In matrix form, this can also be written

$$\begin{pmatrix}\mathbf{c}_1^{\mathrm{T}}\mathbf{A}^{\delta_1 - 1}\mathbf{B}\\ \vdots \\ \mathbf{c}_p^{\mathrm{T}}\mathbf{A}^{\delta_p - 1}\mathbf{B}\end{pmatrix}\mathbf{M}\,\mathbf{y}_r = \begin{pmatrix}k_1 & 0 & \cdots & 0\\ 0 & \ddots & & \vdots\\ \vdots & & \ddots & \vdots\\ 0 & \cdots & \cdots & k_p\end{pmatrix}\mathbf{y}_r\,.$$

Since this equality must hold whatever $\mathbf{y}_r$, it leads with the use of (1.48) to

$$\mathbf{D}^{*}\mathbf{M} = \mathrm{diag}(k_1,\dots,k_p)\,. \tag{1.51}$$

In order to complete the decoupling of (1.49) it is further necessary to express the coefficient of **x** in this equation as a function of y_i and its derivatives. One degree of freedom remains available to do this, the **L** matrix. It is enough then, to let

$$(\mathbf{c}_i^{\mathrm{T}}\mathbf{A}^{\delta_i} - \mathbf{c}_i^{\mathrm{T}}\mathbf{A}^{\delta_i - 1}\mathbf{BL})\mathbf{x} = -(q_{i,\delta_i - 1}\,y_i^{(\delta_i - 1)} + \ldots + q_{i,1}\,\dot{y}_i + q_{i,0}\,y_i)\,, \tag{1.52}$$

and to replace the successive derivatives of y_i by their values derived from (1.47), to see that **L** must then satisfy

$$(\mathbf{c}_i^{\mathrm{T}}\mathbf{A}^{\delta_i} - \mathbf{c}_i^{\mathrm{T}}\mathbf{A}^{\delta_i - 1}\mathbf{BL})\mathbf{x} = -(q_{i,\delta_i - 1}\,\mathbf{c}_i^{\mathrm{T}}\mathbf{A}^{\delta_i - 1}\mathbf{x} + \ldots + q_{i,1}\,\mathbf{c}_i^{\mathrm{T}}\mathbf{A}\mathbf{x} + q_{i,0}\,\mathbf{c}_i^{\mathrm{T}}\mathbf{x})\,,$$

thus also, since this equality must hold for any **x**,

$$\mathbf{c}_i^{\mathrm{T}}\mathbf{A}^{\delta_i-1}\mathbf{B}\mathbf{L} = \mathbf{c}_i^{\mathrm{T}}\mathbf{A}^{\delta_i} + q_{i,\delta_i-1}\,\mathbf{c}_i^{\mathrm{T}}\mathbf{A}^{\delta_i-1} + \ldots + q_{i,1}\,\mathbf{c}_i^{\mathrm{T}}\mathbf{A} + q_{i,0}\,\mathbf{c}_i^{\mathrm{T}}$$
$$= \mathbf{c}_i^{\mathrm{T}} D_i(\mathbf{A}), \quad i = 1,\ldots,p,$$

where the second equal sign results from (1.45). Regrouping these p equations in matrix form with the use of (1.48), we obtain

$$\mathbf{D}^{*}\,\mathbf{L} = \begin{pmatrix} \mathbf{c}_1^{\mathrm{T}} D_1(\mathbf{A}) \\ \vdots \\ \mathbf{c}_p^{\mathrm{T}} D_p(\mathbf{A}) \end{pmatrix}. \tag{1.53}$$

Substituting now (1.52) and (1.50) into (1.49), we find

$$y_i^{(\delta_i)} + q_{i,\,\delta_i-1}\, y_i^{(\delta_i-1)} + \ldots + q_{i,1}\,\dot{y}_i + q_{i,0}\, y_i = k_i\, y_{r,i}\,,$$

equation which confirms that the decoupling is total, and, at the same time, retrieves, by applying the Laplace transform, the expression of the transfer function $G_i(s) = Y_i(s)/Y_{r,i}(s)$ given by (1.43) to (1.45), for $i = 1,\ldots,p$.

The expected control law will thus be obtained by solving (1.51) and (1.53) for **M** and **L**, which will be possible only if $\mathbf{D}^{*}$ is invertible.

Let us summarize these results:

Theorem 1.5. *Consider a plant* $\begin{cases} \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \\ y_i = \mathbf{c}_i^{\mathrm{T}}\mathbf{x}, \; i = 1,\ldots,p \end{cases}$. *It is possible to design a control law* $\mathbf{u} = -\mathbf{L}\mathbf{x} + \mathbf{M}\mathbf{y}_r$ *that decouples the closed-loop system if, and only if,*

$$\det \mathbf{D}^{*} = \det \begin{pmatrix} \mathbf{c}_1^{\mathrm{T}}\mathbf{A}^{\delta_1-1}\mathbf{B} \\ \vdots \\ \mathbf{c}_p^{\mathrm{T}}\mathbf{A}^{\delta_p-1}\mathbf{B} \end{pmatrix} \neq 0\,, \tag{1.54}$$

where δ_i *is the order difference of the output* y_i, *determined by (1.46).*

The parameters of this control law are then given by

$$\mathbf{L} = \mathbf{D}^{*-1} \begin{pmatrix} \mathbf{c}_1^{\mathrm{T}} D_1(\mathbf{A}) \\ \vdots \\ \mathbf{c}_p^{\mathrm{T}} D_p(\mathbf{A}) \end{pmatrix} = \mathbf{D}^{*-1} \begin{pmatrix} \mathbf{c}_1^{\mathrm{T}} \sum\limits_{\nu=0}^{\delta_1} q_{1,\nu}\,\mathbf{A}^{\nu} \\ \vdots \\ \mathbf{c}_p^{\mathrm{T}} \sum\limits_{\nu=0}^{\delta_p} q_{p,\nu}\,\mathbf{A}^{\nu} \end{pmatrix}; \quad \mathbf{M} = \mathbf{D}^{*-1} \begin{pmatrix} k_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & k_p \end{pmatrix}. \tag{1.55}$$

The closed-loop system is then equivalent to the following p distinct SISO systems:

$$Y_i(s) = \frac{k_i}{s^{\delta_i} + q_{i,\delta_i -1}\, s^{\delta_i -1} + \ldots + q_{i,1}\, s + q_{i,0}}\, Y_{r,i}(s), \quad i = 1,\ldots,p\,.$$

The still free parameters $q_{i,\nu}$ *can then be determined separately for each of these systems, e.g. by pole placement, and one can impose additionally* $k_i = q_{i,0}$ *in order to guarantee the static accuracy of each of them.*

1.5.4 Control Implementation

1.5.4.1 Case where $\delta = n$: Maximal Order-difference System

Since we assign in this case $\delta = \delta_1 + \ldots + \delta_p = n$ poles to the various branches of the decoupled system, poles which are at the same time eigenvalues of the latter, it is thus possible here to shift *all* its eigenvalues. An example will illustrate this case in the solved exercises of this chapter.

1.5.4.2 Case where $\delta < n$

The whole set of SISO systems equivalent to the closed-loop system has here a number δ of eigenvalues, lower than that of the closed-loop. It has been shown, [Föl90], [Rop90], that the control law determined by this method assigns then to the closed-loop system zeros in its transfer matrix, which cancel these missing eigenvalues, so that they do not appear any more in the decoupled structure. [Rop90] has even shown that one feasibility criterion of the decoupling by state feedback, equivalent to (1.54), consists in verifying whether the plant possesses exactly $n - \delta$ zeros. For a discussion about internal stability, see Sect. 1.6.5, Remark 1.15.

1.6 Complete Modal Control

The method described in Sect. 1.4 allows shifting only a number *p* of eigenvalues equal to that of the plant inputs, very often lower than the plant order *n* in the case of MIMO systems.

The method which will be described hereafter, and which has been introduced by [Rop82], allows assigning all the closed-loop eigenvalues, therefore its name *complete modal control*, also *parametric controller design*. Unlike the simple modal control, the modal quantities of the closed-loop system, not of the plant, will be used here.

1.6.1 Control Law Derivation (MIMO Case): Roppenecker's Formula

Let us consider again a linear time-invariant system, $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu}$ in the continuous case, or $\mathbf{x}_{k+1} = \mathbf{\Phi x}_k + \mathbf{\Gamma u}_k$ in the discrete case. Assume in the following that

1. the eigenvalues $\lambda_1,\ldots,\lambda_n$ of **A** or of $\mathbf{\Phi}$ are all distinct;
2. a control law $\mathbf{u} = -\mathbf{L}\mathbf{x}$ has been synthesized, which yields the closed-loop eigenvalues $\lambda_{L1},\ldots,\lambda_{Ln}$, also all distinct. The way **L** was obtained, e.g. by pole assignment or even by LQ design as described in Chap. 3, is unimportant.

The case of multiple open-loop eigenvalues has been taken into account in [Rop86], but involves deeper calculations than the ones to be presented here.

The following will deal only with the continuous case, the calculations being easily transposed to the discrete case by merely replacing the **A** and **B** matrices by $\mathbf{\Phi}$ and $\mathbf{\Gamma}$, respectively. The eigenvector $\mathbf{v}_{Li}$ of the closed-loop system matrix $\mathbf{A}-\mathbf{BL}$ and the corresponding eigenvalue λ_{Li} satisfy the equation

$$\left[\lambda_{Li}\mathbf{I} - (\mathbf{A}-\mathbf{BL})\right]\mathbf{v}_{Li} = 0, \quad i = 1,\ldots,n, \tag{1.56}$$

which can be written also

$$(\mathbf{A}-\lambda_{Li}\mathbf{I})\mathbf{v}_{Li} = \mathbf{BLv}_{Li}, \quad i = 1,\ldots,n. \tag{1.57}$$

Let us introduce *n parameter vectors*, of size *p*, defined by

$$\mathbf{p}_i = \mathbf{Lv}_{Li}, \quad i = 1,\ldots,n \ . \tag{1.58}$$

(1.57) becomes then, successively

$$(\mathbf{A}-\lambda_{Li}\mathbf{I})\mathbf{v}_{Li} = \mathbf{Bp}_i, \quad i = 1,\ldots,n, \tag{1.59}$$

$$\mathbf{v}_{Li} = (\mathbf{A}-\lambda_{Li}\mathbf{I})^{-1}\mathbf{Bp}_i, \quad i = 1,\ldots,n. \tag{1.60}$$

The existence of $(\mathbf{A}-\lambda_{Li}\mathbf{I})^{-1}$ is guaranteed if we assume temporarily that the λ_{Li} haven been chosen different from the λ_i, thus that all $\mathbf{A}$ eigenvalues have been shifted. This restriction will be removed in Remark 1.12 below.

Let us gather the $\mathbf{p}_i$ and $\mathbf{v}_{Li}$ vectors in two matrices:

$$\begin{pmatrix}\mathbf{p}_1 & \cdots & \mathbf{p}_n\end{pmatrix}=\mathbf{L}\begin{pmatrix}\mathbf{v}_{L1} & \cdots & \mathbf{v}_{Ln}\end{pmatrix}. \tag{1.61}$$

Since the λ_{Li} are all different, the eigenvectors $\mathbf{v}_{Li}$ are linearly independent.

Let us go now the reverse way, by choosing arbitrarily the λ_{Li} 's *and the* $\mathbf{p}_i$'s, $\mathbf{L}$ remaining to be determined. The two previous equations lead thus to the following theorem:

Theorem 1.6. *Consider a MIMO continuous-time system* $\dot{\mathbf{x}}=\mathbf{Ax}+\mathbf{Bu}$*, assumed controllable and having constant matrices* $\mathbf{A}$ *(* $n\times n$ *) and* $\mathbf{B}$ *(* $n\times p$ *). Let us choose arbitrarily n values* $\lambda_{L1},\ldots,\lambda_{Ln}$ *and n vectors* $\mathbf{p}_1,\ldots,\mathbf{p}_n$ *of size p, with the unique constraint that the vectors*

$$\mathbf{v}_{Li}=(\mathbf{A}-\lambda_{Li}\mathbf{I})^{-1}\mathbf{B}\mathbf{p}_i,\quad i=1,\ldots,n, \tag{1.62}$$

resulting from these two choices are linearly independent. The closed-loop system obtained by a state-feedback control $\mathbf{u}=-\mathbf{Lx}$ *with*

$$\begin{aligned}\mathbf{L}&=\begin{pmatrix}\mathbf{p}_1 & \cdots & \mathbf{p}_n\end{pmatrix}\begin{pmatrix}\mathbf{v}_{L1} & \cdots & \mathbf{v}_{Ln}\end{pmatrix}^{-1}\\&=\begin{pmatrix}\mathbf{p}_1 & \cdots & \mathbf{p}_n\end{pmatrix}\begin{pmatrix}(\mathbf{A}-\lambda_{L1}\mathbf{I})^{-1}\mathbf{B}\mathbf{p}_1 & \cdots & (\mathbf{A}-\lambda_{Ln}\mathbf{I})^{-1}\mathbf{B}\mathbf{p}_n\end{pmatrix}^{-1}\end{aligned} \tag{1.63}$$

has the eigenvalues $\lambda_{L1},\ldots,\lambda_{Ln}$ *and the eigenvectors* $\mathbf{v}_{L1},\ldots,\mathbf{v}_{Ln}$ *given by (1.62).*

Relation (1.63) is known as the *Roppenecker's formula*. It calls for several remarks.

Remark 1.9. If one of the chosen eigenvalues, λ_{Li}, is complex, its complex conjugate $\overline{\lambda}_{Li}$ must also be selected. Failing to do so would cause the $\mathbf{L}$ matrix to contain complex elements. For the same reason, two complex-conjugated parameter vectors must be chosen in association with this pair of eigenvalues.

Remark 1.10. The closed-loop eigenvectors $\mathbf{v}_{Li}$ being linearly independent, they build a basis of the state space, in which the closed-loop state vector of the free response to an initial state $\mathbf{x}_0$ is expressed by $\mathbf{x}=\sum_{i=1}^{n}x_i^*\,\mathbf{v}_{Li}$, with $x_i^*=e^{\lambda_{Li}t}\,\mathbf{w}_{Li}^{\mathrm{T}}\,\mathbf{x}_0$

(see (1.66)). Left multiplying this equation by **L**, we obtain for the opposite control vector

$$-\mathbf{u} = \mathbf{L}\,\mathbf{x} = \sum_{i=1}^{n} x_i^* \,\mathbf{L}\,\mathbf{v}_{Li} = \sum_{i=1}^{n} x_i^* \,\mathbf{p}_i \ ,$$

according to (1.58). The comparison of these two equations shows that the parameter vectors form a basis of the control vectors subspace. Furthermore, the control vector $\mathbf{u}(t)$, resulting from the application of the complete modal control law to the state vector $\mathbf{x}(t)$ at a given time, has in this basis the *same components* as the state vector in the basis of the closed-loop eigenvectors.

Remark 1.11. Parameter-vector invariance in a regular transformation.

Let us submit the state vector to a change of basis, defined by a regular transformation matrix **T**: $\mathbf{x} = \mathbf{T}\mathbf{z}$. The state equations become (Sect. A.7.1 of Appendix A)

$$\dot{\mathbf{z}} = \hat{\mathbf{A}}\,\mathbf{z} + \hat{\mathbf{B}}\,\mathbf{u}, \quad \hat{\mathbf{A}} = \mathbf{T}^{-1}\mathbf{A}\mathbf{T}, \quad \hat{\mathbf{B}} = \mathbf{T}^{-1}\mathbf{B}\,,$$

and the control law writes

$$\mathbf{u} = -\mathbf{L}\mathbf{x} = -\hat{\mathbf{L}}\,\mathbf{z}, \ \text{with} \ \hat{\mathbf{L}} = \mathbf{L}\mathbf{T}\,. \tag{1.64}$$

The eigenvalues of a matrix remaining unchanged in a change of basis, the closed-loop eigenvectors satisfy, in the new coordinates, the following equation:

$$\left[\lambda_{Li}\,\mathbf{I} - (\hat{\mathbf{A}} - \hat{\mathbf{B}}\hat{\mathbf{L}})\right]\hat{\mathbf{v}}_{Li} = 0\,,$$

which can also be written

$$\left[\lambda_{Li}\,\mathbf{T}^{-1}\mathbf{T} - (\mathbf{T}^{-1}\mathbf{A}\mathbf{T} - \mathbf{T}^{-1}\mathbf{B}\mathbf{L}\mathbf{T})\right]\hat{\mathbf{v}}_{Li} = 0\,,$$

thus, by left multiplication with **T,**

$$\left[\lambda_{Li} - (\mathbf{A} - \mathbf{B}\mathbf{L})\right]\mathbf{T}\,\hat{\mathbf{v}}_{Li} = 0\,.$$

The comparison of the last equation with (1.56) shows that

$$\hat{\mathbf{v}}_{Li} = \mathbf{T}^{-1}\,\mathbf{v}_{Li}\,. \tag{1.65}$$

The eigenvectors are transformed thus the same way as the state vector coordinates. For the parameter vectors in the new coordinates, however, their definition (1.58) combined with (1.65) leads to

$$\hat{\mathbf{p}}_i = \hat{\mathbf{L}}\hat{\mathbf{v}}_{Li} = \mathbf{L}\mathbf{T}\mathbf{T}^{-1}\mathbf{v}_{Li} = \mathbf{L}\,\mathbf{v}_{Li} = \mathbf{p}_i\,.$$

In conclusion, the parameter vectors are invariant in a regular transformation, as are the eigenvalues, whereas the *eigenvectors* and the state-feedback matrix are not (see (1.65) and (1.64)). They reflect thus, in the same way as the eigenvalues, an intrinsic property of the control system, and are also called for this reason *invariant* parameter vectors [Rop81], [Rop90].

Remark 1.12. While writing (1.60), we have assumed that all initial eigenvalues of the plant were shifted, which guaranteed that $\det(\mathbf{A}-\lambda_{Li}\,\mathbf{I}) \neq 0, \forall i = 1,\ldots,n$. This restriction can be removed easily.

Suppose, indeed, that we would have wanted to keep a given initial eigenvalue λ_i. It is then logic to associate with the choice $\lambda_{Li} = \lambda_i$ the requirement that the eigenvector corresponding to this eigenvalue remains itself also unchanged: $\mathbf{v}_{Li} = \mathbf{v}_i$. Then, according to (1.59),

$$(\mathbf{A}-\lambda_i\,\mathbf{I})\mathbf{v}_i = \mathbf{B}\mathbf{p}_i\,,$$

thus

$$0 = \mathbf{B}\mathbf{p}_i\,,$$

since $\mathbf{v}_i$ is the eigenvector associated with λ_i. As a result, $\mathbf{p}_i = 0$, since **B** is of full rank. This leads to the

Complementary Rule to Theorem 1.6. *If an eigenvalue of the initial plant should not be shifted* $(\lambda_{Li} = \lambda_i)$, *let* $\mathbf{p}_i = 0$ *in the formula (1.63) which yields* **L**, *and replace there*

$$\mathbf{v}_{Li} = (\mathbf{A}-\lambda_{Li}\,\mathbf{I})^{-1}\,\mathbf{B}\mathbf{p}_i$$

by the eigenvector $\mathbf{v}_i$ *of the plant.*

Remark 1.13: choice of the parameter vectors. In the design of a state feedback for a MIMO system, the degrees of freedom, which are left once the closed-loop eigenvalues have been chosen, are expressed in explicit form by the parameter vectors. These vectors are indeed determined, like the eigenvectors, apart from one multiplicative factor. They represent thus a total of $n(p-1)$ parameters, which, added to the n prescribed eigenvalues, make up for the total number np of design parameters required to determine the *np* elements of the **L** matrix in the MIMO case. It is obvious that these additional degrees of freedom do not exist in the SISO case.

As mentioned previously, the parameter vectors are chosen arbitrarily, and this choice can be completely at random. This way of doing yields the highest probability that the eigenvectors resulting from (1.62) be, as required, linearly independent.

It is however possible to refine this choice, in order to fulfill some additional requirements for the closed-loop, e.g. specifications of structural nature. One of these possibilities will be the subject of the next section.

1.6.2 Targeted Choice of the Parameter Vectors and the Eigenvectors

1.6.2.1 Influence of the Eigenvectors on the Closed-loop System Dynamics

The dynamic response of the closed-loop system state to non-vanishing initial conditions $\mathbf{x}_0$ and to an input stimulation, the reference $\mathbf{y}_r(t)$, is given, according to (1.3) and to (A.18) of Appendix A, where the matrix **A** is replaced by $\mathbf{A}-\mathbf{BL}$ and **u** by $\mathbf{M}\mathbf{y}_r(t)$, by the equation

$$\mathbf{x}(t) = e^{(\mathbf{A}-\mathbf{BL})t}\,\mathbf{x}_0 + \int_0^t e^{(\mathbf{A}-\mathbf{BL})(t-\tau)}\,\mathbf{B}\,\mathbf{M}\mathbf{y}_r(\tau)\,d\tau\,.$$

The closed-loop equivalent to (A.45) can therefore be written as

$$\mathbf{x}(t) = \sum_{i=1}^{n} \underbrace{\mathbf{v}_{Li}\,e^{\lambda_{Li}t}\mathbf{w}_{Li}^{\mathrm{T}}}_{\text{closed-loop eigenmode}}\;\mathbf{x}_0 + \sum_{i=1}^{n}\mathbf{v}_{Li}\int_0^t e^{\lambda_{Li}(t-\tau)}\mathbf{w}_{Li}^{\mathrm{T}}\,\mathbf{B}\mathbf{M}\mathbf{y}_r(\tau)\,d\tau\,. \quad (1.66)$$

In this equation, $\mathbf{v}_{Li}$ and $\mathbf{w}_{Li}$ represent respectively the eigenvector and the left eigenvector (see Appendix A, Sect. A.7.4) associated with the closed-loop eigenvalue λ_{Li}. The set of these eigenvalues and eigenvectors constitutes the *eigenstructure* of the closed-loop system.

By transposing the discussion following (A.45) to the closed-loop situation, it must therefore be possible to decouple certain modes $e^{\lambda_{Li}t}$ from certain states, thus from certain outputs, by choosing appropriately the associated eigenvectors $\mathbf{v}_{Li}$ [Duc01].

In a symmetric way, it must be possible also to decouple certain transient modes $e^{\lambda_{Lj} t}$ from certain components of the initial state $\mathbf{x}_0$ or from certain components of the reference $\mathbf{y}_r(t)$, by choosing appropriately the left eigenvectors $\mathbf{w}_{Lj}^{\mathrm{T}}$ of the closed-loop system. The interested reader may find more details about this topic in specialized works, such as e.g. [LiKM94], [LKMA94]. We will limit ourselves here to the choice of the closed-loop (right) eigenvectors.

1.6.2.2 Coupled Choice of the Parameter Vectors and the Eigenvectors

Let us start again from (1.59), which can also be written

$$\begin{pmatrix} \mathbf{A} - \lambda_{Li}\,\mathbf{I}_n & -\mathbf{B} \end{pmatrix} \begin{pmatrix} \mathbf{v}_{Li} \\ \mathbf{p}_i \end{pmatrix} = 0, \quad i = 1, \ldots, n. \tag{1.67}$$

If all prescribed values λ_{Li} are different from the eigenvalues of $\mathbf{A}$, the matrix $\begin{pmatrix} \mathbf{A} - \lambda_{Li}\,\mathbf{I}_n & -\mathbf{B} \end{pmatrix}$ is of rank n and (1.67) represents a system of n linear equations in $(n+p)$ unknowns. It is thus possible to choose arbitrarily the value of p of them. The resulting system of equations has then one unique solution for the n other unknowns.

The arbitrary choice of p unknowns can be spread indifferently over the components of $\mathbf{v}_{Li}$ and those of $\mathbf{p}_i$. Thus, e.g., the choice of making the j-th component of $\mathbf{v}_{Li}$ vanish will cause the mode $e^{\lambda_{Li} t}$ to be absent from the j-th component of $\mathbf{x}$. A real example developed in the exercises of this chapter will illustrate this approach.

1.6.3 Single-input Systems (p = 1)

The parameter vectors are here scalars $p_i \neq 0$, so as to guarantee the independence of the eigenvectors $\mathbf{v}_{Li}$. The matrix in the second factor on the right side of (1.63) has then the following expression, with $\mathbf{B} = \mathbf{b}$:

$$\begin{aligned} &\begin{pmatrix} (\mathbf{A} - \lambda_{L1}\,\mathbf{I})^{-1}\mathbf{b}\,p_1 & \cdots & (\mathbf{A} - \lambda_{Ln}\,\mathbf{I})^{-1}\mathbf{b}\,p_n \end{pmatrix} \\ &\qquad = \begin{pmatrix} (\mathbf{A} - \lambda_{L1}\,\mathbf{I})^{-1}\mathbf{b} & \cdots & (\mathbf{A} - \lambda_{Ln}\,\mathbf{I})^{-1}\mathbf{b} \end{pmatrix} \begin{pmatrix} p_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & p_n \end{pmatrix}. \end{aligned}$$

This formula yields then the state feedback matrix

$$\mathbf{L} = \boldsymbol{\ell}^{\mathrm{T}} = \begin{pmatrix} p_1 & \cdots & p_n \end{pmatrix} \begin{pmatrix} \dfrac{1}{p_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \dfrac{1}{p_n} \end{pmatrix} \Big((\mathbf{A} - \lambda_{L1}\mathbf{I})^{-1}\mathbf{b} \quad \cdots \quad (\mathbf{A} - \lambda_{Ln}\mathbf{I})^{-1}\mathbf{b} \Big)^{-1}$$

$$= \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix} \Big((\mathbf{A} - \lambda_{L1}\mathbf{I})^{-1}\mathbf{b} \quad \cdots \quad (\mathbf{A} - \lambda_{Ln}\mathbf{I})^{-1}\mathbf{b} \Big)^{-1} .$$

Notice that the parameter vectors have disappeared from the expression of $\boldsymbol{\ell}^{\mathrm{T}}$, which thus depends only on the eigenvalues λ_{Li}. This is not a surprise, since in the single input case there is no additional degree of freedom.

The equivalence with the Ackermann's formula is proved by comparison of the closed-loop characteristic polynomials in the two approaches [Rop82].

1.6.4 Simple Modal Control Contained in Complete Modal Control

As it was seen in Sect. 1.4.2.1, only a number p of eigenvalues equal to the number of the plant inputs are shifted in the approach of simple modal control. With the notations of that section, let us choose then the matrix of parameter vectors in the form

$$\mathbf{P} = \Big(\hat{\mathbf{B}}_p^{-1}(\boldsymbol{\Lambda}_p - \boldsymbol{\Lambda}_{Lp}) \quad \mathbf{0} \Big), \tag{1.68}$$

which amounts to taking for the first p parameter vectors

$$\mathbf{p}_i = \hat{\mathbf{B}}_p^{-1}\mathbf{e}_i(\lambda_i - \lambda_{Li}), \quad i = 1, \ldots, p ,$$

where $\mathbf{e}_i = \begin{pmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{pmatrix}^{\mathrm{T}}$, the non-vanishing coefficient being at the i-th place, and

$$\mathbf{p}_i = 0, \quad i = p+1, \ldots, n ,$$

for the $(n-p)$ remaining ones. By substitution of (1.68) into (1.63) the state feedback matrix becomes

$$\mathbf{L} = \hat{\mathbf{B}}_p^{-1}(\mathbf{\Lambda}_p - \mathbf{\Lambda}_{Lp})\begin{pmatrix}\mathbf{v}_{L1} & \cdots & \mathbf{v}_{Ln}\end{pmatrix}^{-1}.$$

A more in-depth calculation [Rop81], which involves the relation existing between the closed-loop eigenvectors, $\mathbf{v}_{Li}$, and those of the plant, $\mathbf{v}_i$, as well as the orthogonality of the left and the right eigenvectors of a matrix, shows then that the above expression reduces to

$$\mathbf{L} = \hat{\mathbf{B}}_p^{-1}(\mathbf{\Lambda}_p - \mathbf{\Lambda}_{Lp})\begin{pmatrix}\mathbf{w}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{w}_p^{\mathrm{T}}\end{pmatrix},$$

which is nothing else but (1.40).

1.6.5 Input – Output Decoupling Control by Complete Modal Design

As in the case of the Falb-Wolovich method exposed in Sect. 1.5, this synthesis is possible only if $p = q$. Starting from these formulae, [Rop90] has shown that, for a system controlled by such a control law, if an eigenvector is associated with a desired eigenvalue λ_{Li} which appears as one of the poles $p_{j,k}$ of the diagonal element $G_j(s)$ according to (1.43)-(1.45), it satisfies the relation

$$\mathbf{C}\mathbf{v}_{Li} = \mathbf{e}_j, \text{ if } \lambda_{Li} = p_{j,k}, \quad k \in \{1,\ldots,\delta_k\}, \tag{1.69}$$

where $\mathbf{e}_j = \begin{pmatrix}0 & \cdots & 0 & 1 & 0 & \cdots & 0\end{pmatrix}^{\mathrm{T}}$ is a unit vector of size q, the only non zero element of which is the one at the j-th place. If on the contrary it is associated with an eigenvalue which is not equal to any of these poles $p_{j,k}$, it satisfies

$$\mathbf{C}\mathbf{v}_{Li} = 0, \text{ if } \lambda_{Li} \neq p_{j,k}. \tag{1.70}$$

Introducing with (1.60) the invariant parameter vector $\mathbf{p}_i$ associated with this eigenvalue yields

$$\mathbf{C}(\mathbf{A} - \lambda_{Li}\mathbf{I})^{-1}\mathbf{B}\mathbf{p}_i = \mathbf{e}_j,$$

thus also

$$-\mathbf{G}(\lambda_{Li})\mathbf{p}_i = \mathbf{e}_j\,,$$

where $\mathbf{G}(s)$ is the transfer matrix of the open-loop plant.

In summary, the steps of the synthesis are the following:

1. Prescribe for the closed-loop system $n-\delta$ eigenvalues λ_{Li} equal to the $n-\delta$ zeros of the plant to be decoupled, with corresponding eigenvector $\mathbf{v}_{Li} = \mathbf{x}_Z$ and parameter vector $\mathbf{p}_i = -\mathbf{u}_Z$, where $\mathbf{x}_Z$ and $\mathbf{u}_Z$ are the *state direction* and *control direction associated with these zeros* (Appendix A, Sect. A.6);
2. The δ remaining closed-loop eigenvalues can be chosen arbitrarily, the parameter vectors associated with these λ_{Li} being given by

$$\mathbf{p}_i = -\mathbf{G}^{-1}(\lambda_{Li})\mathbf{e}_j\,. \tag{1.71}$$

The choice of these eigenvalues must occur such that a number δ_j of them be assigned as poles $p_{j,k}$ of the *j*-th transfer function $G_j(s)$, with $k = 1,\ldots,\delta_j$;

3. The matrix **L** is then calculated according to the procedure of the complete modal synthesis (Theorem 1.6, Equation (1.63)) and the gain compensation (feedforward) matrix **M** is determined, either by (1.54) and (1.55) or their discrete equivalent, or by the general case relation (1.8) or (1.16).

Remark 1.14. By grouping together equations (1.59) and (1.70), we obtain the homogeneous linear equation system

$$\begin{pmatrix} \mathbf{A}-\lambda_{Li}\mathbf{I} & \mathbf{B} \\ \mathbf{C} & \mathbf{0} \end{pmatrix}\begin{pmatrix} \mathbf{v}_{Li} \\ -\mathbf{p}_i \end{pmatrix} = 0\,,$$

the solution pairs $(\lambda_{Li}, \mathbf{v}_{Li})$ of which are precisely the plant zeros and their associated directions. This confirms the choices which have been imposed in the first step of the design procedure.

Remark 1.15. Case where the plant has "unstable" zeros.

It would be possible a priori to proceed as previously, in Step 1, by placing on these zeros closed-loop eigenvalues in order to cancel them. A complete input-output decoupling would indeed result for the closed-loop system, with a diagonal transfer matrix, but we would have created this way an *unstable internal mode* for the closed-loop, even though it would be unobservable from the output.

The complete modal design provides a more realistic solution, by avoiding this compensation, even if it means giving up complete decoupling. It is enough then to place closed-loop eigenvalues in Step 1 only on "*stable*" plant zeros, and to join the eigenvalues which have become free this way with the other ones, which are used in Step 2 of the approach.

Suppose e.g. that such left-aside eigenvalues have been assigned to the j-th output. The closed-loop transfer matrix will have the following shape (with $p = q$):

$$\mathbf{G}_{CL}(s) = \begin{pmatrix} G_1(s) & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & & & \vdots \\ G_{j1}(s) & \cdots & G_{jj}(s) & \cdots & G_{jp}(s) \\ \vdots & & & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & G_p(s) \end{pmatrix},$$

to compare with (1.42). This approach is called *partial input-output decoupling*.

Conclusion: a system can thus be put in a totally decoupled shape, provided it has neither any zero with positive real part (or exterior to the unit circle, for discrete systems), nor any zero which would be common to all the elements of a same row of the plant transfer matrix $\mathbf{G}_{OL}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$, this last situation being rarely encountered in practice (see Sect. 1.5.2).

1.6.6 Disturbance – Output Decoupling by Complete Modal Design

In the previous section, it has been shown that the complete modal design makes it possible to realize a decoupling control, which decouples the input-output links.

It is also possible, in certain circumstances, to eliminate the influence of constant disturbances on the outputs by means of the complete modal design, thus to create a *disturbance-output* decoupling, [Rop90].

Let us consider a continuous-time system with the same number of inputs and outputs, $p = q$, described by

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{E}\mathbf{v} \\ \mathbf{y} = \mathbf{C}\mathbf{x} \end{cases},$$

where $\mathbf{v}$ is a constant disturbance vector of size m, influencing the state through an $(n \times m)$ matrix $\mathbf{E}$, as illustrated in the diagram of Fig. 1.8, which will bee seen in Sect. 1.8.1.4.

Let us assume that it is equipped with a control law of type (1.2): $\mathbf{u} = -\mathbf{L}\,\mathbf{x} + \mathbf{M}\,\mathbf{y}_r$. The closed-loop system is then represented by

$$\begin{cases} \dot{\mathbf{x}} = (\mathbf{A} - \mathbf{B}\mathbf{L})\mathbf{x} + \mathbf{B}\,\mathbf{M}\,\mathbf{y}_r + \mathbf{E}\,\mathbf{v} \\ \mathbf{y} = \mathbf{C}\mathbf{x} \end{cases}.$$

The transfer matrix between the disturbance input **v** and the output **y**, thus at vanishing reference input $\mathbf{y}_r$, is given by

$$\mathbf{G}_v(s) = \mathbf{C}\big[s\mathbf{I} - (\mathbf{A} - \mathbf{BL})\big]^{-1}\mathbf{E}\,. \tag{1.72}$$

Let us go over to modal representation, with the change of variable $\mathbf{x} = \mathbf{T}\mathbf{x}^*$, $\mathbf{T}$ being the modal matrix. If the eigenvalues prescribed for the closed-loop system are assumed all distinct, its transfer matrix will have in this basis the diagonal form $\mathbf{A} - \mathbf{BL} = diag(\lambda_{L1}, \dots, \lambda_{Ln})$, and (1.72) becomes, by returning to the initial representation,

$$\mathbf{G}_v(s) = \mathbf{CT}\begin{pmatrix} s - \lambda_{L1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s - \lambda_{Ln} \end{pmatrix}^{-1} \mathbf{T}^{-1}\mathbf{E}\,,$$

which, by introducing the eigenvectors $\mathbf{v}_{Li}$ and the left eigenvectors $\mathbf{w}_{Li}$, the transposes of which are the rows of $\mathbf{T}^{-1}$ (see Appendix A, Sect. A.7.4), yields the final formula:

$$\mathbf{G}_v(s) = \sum_{i=1}^{n} \frac{\mathbf{C}\,\mathbf{v}_{Li}\,\mathbf{w}_{Li}^{\mathrm{T}}\,\mathbf{E}}{s - \lambda_{Li}}\,.$$

The desired decoupling will be obtained if one succeeds making that $\mathbf{G}_v(s) = 0$, i.e. that $\mathbf{C}\,\mathbf{v}_{Li}\,\mathbf{w}_{Li}^{\mathrm{T}}\,\mathbf{E} = 0,\ \forall i = 1, \dots, n$, which leads by summation on i to the necessary condition

$$\mathbf{CE} = \mathbf{0}\,. \tag{1.73}$$

For that condition to be satisfied, it will suffice in fact that

$$\begin{cases} \mathbf{C}\,\mathbf{v}_{Li} = 0 & \text{for} \quad i = 1, \dots, k \\ \mathbf{w}_{Li}^{\mathrm{T}}\,\mathbf{E} = 0 & \text{for} \quad i = k+1, \dots, n, \end{cases}$$

where the eigenvalues and eigenvectors may have eventually been renumbered. Grouping together these two equations yields

$$\begin{pmatrix} \mathbf{v}_{L1} & \cdots & \mathbf{v}_{Ln} \end{pmatrix}^{-1}\mathbf{E} = \begin{pmatrix} \mathbf{N} \\ \hdashline \mathbf{0} \end{pmatrix} \begin{matrix} \}\,k \\ \}\,n-k \end{matrix}$$

where **N** is some $(k \times m)$ matrix, from where follows that

$$\mathbf{E} = \begin{pmatrix} \mathbf{v}_{L1} & \cdots & \mathbf{v}_{Lk} \end{pmatrix} \mathbf{N} . \tag{1.74}$$

This relation shows that the disturbance-output decoupling is possible if, and only if, each column of **E** can be written as a linear combination of the prescribed eigenvectors for which $\mathbf{C}\mathbf{v}_{Li} = 0$ holds, in other terms, if, and only if, only unobservable modes are excited by these disturbances.

Now, according to Remark 1.14, such an eigenvector corresponds to an eigenvalue which has been placed on a plant zero, and its direction is equal to that of the state associated with this zero.

The design procedure, according to [Rop90], is then the following:

1. check whether the necessary condition (1.73) for the existence of such a solution is fulfilled;
2. determine then the plant state-vector directions associated with a zero from which the columns of the disturbance input matrix **E** are built, and choose the same number of eigenvalues and of parameter vectors, equal respectively to these zeros and to the control vector directions associated with these zeros, with sign reversal;
3. the remaining parameters of the complete modal design are chosen according to other considerations.

An application example of this approach will be described in an exercise at the end of this chapter.

1.7 General Formula for MIMO Design

A more general expression to calculate the state feedback matrix for MIMO systems has been established by [BeOs87]. It will be given here only in the continuous case, since the discrete case is again derived from it trivially by substitution of the matrices $\mathbf{\Phi}$ and $\mathbf{\Gamma}$ to the matrices **A** and **B**.

1.7.1 MIMO Control Law (Becker-Ostertag)

Theorem 1.7. *Consider a linear time-invariant system,* $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$*, supposed controllable. The matrix* $\mathbf{L}$ *of the state-feedback control law* $\mathbf{u} = -\mathbf{L}\mathbf{x}$ *which shifts the eigenvalues of this system towards desired values* λ_{Li}*,* $i = 1, \ldots, n$*, is given by*

$$\mathbf{L} = \mathbf{P}\mathbf{D}^{-1} q_0(\mathbf{A}) \ , \tag{1.75}$$

where

$$\mathbf{D}=\mathbf{Q}_C\begin{pmatrix}\mathbf{P}\,q_1(\mathbf{A}_{can})\\ \mathbf{P}\,q_2(\mathbf{A}_{can})\\ \vdots\\ \mathbf{P}\,q_n(\mathbf{A}_{can})\end{pmatrix},\ \text{with}\ \mathbf{Q}_C=\begin{pmatrix}\mathbf{B} & \mathbf{AB} & \cdots & \mathbf{A}^{n-1}\mathbf{B}\end{pmatrix},\tag{1.76}$$

and

$$\begin{cases}q_0(\lambda)=\lambda^n+a_{n-1}\lambda^{n-1}+\ldots+a_1\lambda+a_0=(\lambda-\lambda_{L1})\ldots(\lambda-\lambda_{Ln})\\ q_1(\lambda)=\lambda^{n-1}+a_{n-1}\lambda^{n-2}+\ldots+a_2\lambda+a_1\\ \qquad\vdots\\ q_i(\lambda)=\lambda^{n-i}+a_{n-1}\lambda^{n-i-1}+\ldots+a_i\\ \qquad\vdots\\ q_n(\lambda)=1\end{cases}\tag{1.77}$$

and where $\mathbf{A}_{can}$ *is the closed-loop system matrix, chosen in one of the known canonical forms, e.g. the controllability form, and having the desired eigenvalues* λ_{Li}, $i=1,\ldots,n$. *The elements of the* $(p\times n)$ *matrix* $\mathbf{P}$ *are the parameters which can be chosen arbitrarily in a MIMO pole-placement design, with the only restriction that* $\mathbf{D}$ *in* (1.76) *be invertible.*

Proof. By means of a regular transformation $\mathbf{x}=\mathbf{V}\,\mathbf{z}$, the closed-loop system matrix

$$\mathbf{A}_{CL}=\mathbf{A}-\mathbf{B}\,\mathbf{L}\tag{1.78}$$

is transformed into a canonical form $\mathbf{A}_{can}$, which is desired to have the prescribed eigenvalues λ_{Li}.

Right multiplying (1.78) by $\mathbf{V}$ and letting

$$\mathbf{P}=\mathbf{LV}\tag{1.79}$$

yields

$$\mathbf{A}_{CL}\mathbf{V}=\mathbf{AV}-\mathbf{BLV}=\mathbf{AV}-\mathbf{B}\,\mathbf{P}\,.$$

Since $\mathbf{A}_{can}=\mathbf{V}^{-1}\mathbf{A}_{CL}\mathbf{V}$ (equation (A.30), Appendix A), it follows from the previous equation that

$$\mathbf{AV}-\mathbf{V}\,\mathbf{A}_{can}=\mathbf{B}\,\mathbf{P}\,.\tag{1.80}$$

The transposition of this equation,

$$\mathbf{V}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}} - \mathbf{A}_{can}^{\mathrm{T}}\mathbf{V}^{\mathrm{T}} = \mathbf{P}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\,,$$

makes it appear formally identical to the equation

$$\mathbf{TA} - \mathbf{FT} = \mathbf{GC}\,, \tag{1.81}$$

which is encountered in the observer theory (see Sect. 2.2.4.6), by introducing the following duality:

$$\begin{array}{rcl} \textit{Observer Theory} & \rightarrow & \textit{Control Theory} \\ \mathbf{T} & \rightarrow & \mathbf{V}^{\mathrm{T}} \\ \mathbf{A} & \rightarrow & \mathbf{A}^{\mathrm{T}} \\ \mathbf{F} & \rightarrow & \mathbf{A}_c^{\mathrm{T}} \\ \mathbf{G} & \rightarrow & \mathbf{P}^{\mathrm{T}} \\ \mathbf{C} & \rightarrow & \mathbf{B}^{\mathrm{T}} \end{array} \tag{1.82}$$

This remarkable duality between control and observation (or estimation) will be reviewed in more details in Chapter 2.

In the case of Equation (1.81), where the matrices $\mathbf{A}$ $(n \times n)$, $\mathbf{F}$ $(r \times r)$, $\mathbf{G}$ $(r \times q)$ and $\mathbf{C}$ $(q \times n)$ are supposed known and where the unknown is the matrix $\mathbf{T}$ $(r \times n)$, [CeBa84] has shown, building on the works of [BoYo68], that, if the eigenvalues of **F** are all different from those of **A**, this equation has a unique solution **T**, given by

$$\mathbf{T} = \begin{pmatrix} \mathbf{G} & \mathbf{FG} & \cdots & \mathbf{F}^{r-1}\mathbf{G} \end{pmatrix} \begin{pmatrix} \mathbf{C}q_1(\mathbf{A}) \\ \mathbf{C}q_2(\mathbf{A}) \\ \vdots \\ \mathbf{C}q_r(\mathbf{A}) \end{pmatrix} q_0^{-1}(\mathbf{A}),$$

where the polynomials $q_i(\lambda)$ are given by (1.77) by letting $n = r$. The application of the previous duality to this result, with $r = n$, yields

$$\mathbf{V}^{\mathrm{T}} = \begin{pmatrix} \mathbf{P}^{\mathrm{T}} & \mathbf{A}_{can}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}} & \cdots & (\mathbf{A}_{can}^{\mathrm{T}})^{n-1}\mathbf{P}^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} \mathbf{B}^{\mathrm{T}}\, q_1(\mathbf{A}^{\mathrm{T}}) \\ \mathbf{B}^{\mathrm{T}}\, q_2(\mathbf{A}^{\mathrm{T}}) \\ \vdots \\ \mathbf{B}^{\mathrm{T}}\, q_n(\mathbf{A}^{\mathrm{T}}) \end{pmatrix} q_0^{-1}(\mathbf{A}^{\mathrm{T}})\,.$$

By transposition, this writes

$$\mathbf{V} = q_0^{-1}(\mathbf{A})\,\mathbf{D}\,, \tag{1.83}$$

where

$$\mathbf{D} = \begin{pmatrix} q_1(\mathbf{A})\mathbf{B} & q_2(\mathbf{A})\mathbf{B} & \cdots & q_n(\mathbf{A})\mathbf{B} \end{pmatrix} \begin{pmatrix} \mathbf{P} \\ \mathbf{P}\mathbf{A}_{can} \\ \vdots \\ \mathbf{P}\mathbf{A}_{can}^{n-1} \end{pmatrix}.$$

By developing the expression of **D,**

$$\mathbf{D} = q_1(\mathbf{A})\mathbf{B}\mathbf{P} + q_2(\mathbf{A})\mathbf{B}\mathbf{P}\mathbf{A}_{can} + \ \ldots\ + q_{n-1}(\mathbf{A})\mathbf{B}\mathbf{P}\mathbf{A}_{can}^{n-2} + q_n(\mathbf{A})\mathbf{B}\mathbf{P}\mathbf{A}_{can}^{n-1}\,,$$

then expressing the polynomials $q_i(\mathbf{A})$ with the use of (1.77), and finally grouping together the terms of same power of **A**, it is readily seen that

$$\begin{aligned}\mathbf{D} &= \mathbf{B}\mathbf{P}q_1(\mathbf{A}_{can}) + \mathbf{A}\mathbf{B}\mathbf{P}q_2(\mathbf{A}_{can}) + \ \ldots\ + \mathbf{A}^{n-2}\mathbf{B}\mathbf{P}q_{n-1}(\mathbf{A}_{can}) + \mathbf{A}^{n-1}\mathbf{B}\mathbf{P}q_n(\mathbf{A}_{can}) \\ &= \begin{pmatrix} \mathbf{B} & \mathbf{A}\mathbf{B} & \cdots & \mathbf{A}^{n-1}\mathbf{B} \end{pmatrix} \begin{pmatrix} \mathbf{P}q_1(\mathbf{A}_{can}) \\ \mathbf{P}q_2(\mathbf{A}_{can}) \\ \vdots \\ \mathbf{P}q_n(\mathbf{A}_{can}) \end{pmatrix} = \mathbf{Q}_C \begin{pmatrix} \mathbf{P}q_1(\mathbf{A}_{can}) \\ \mathbf{P}q_2(\mathbf{A}_{can}) \\ \vdots \\ \mathbf{P}q_n(\mathbf{A}_{can}) \end{pmatrix}.\end{aligned}$$

This expression is nothing else but (1.76). Finally, (1.79) and (1.83) yield

$$\mathbf{L} = \mathbf{P}\mathbf{V}^{-1} = \mathbf{P}\mathbf{D}^{-1}q_0(\mathbf{A})\,,$$

which proves (1.75).

Remark 1.16. Similarities with the complete modal design.

Though the present approach is entirely different, a strong similarity is noticeable with the expressions of the complete modal control design. Compare in particular relations (1.79) and (1.58): the parameter matrix **P** plays here the same role as the matrix $\mathbf{P} = \begin{pmatrix} \mathbf{p}_1 & \cdots & \mathbf{p}_n \end{pmatrix}$ of the parameter vectors introduced by Roppenecker.

Here also, the matrix **P** is invariant in a similitude transformation. Furthermore, (1.79) brings about that $\mathbf{u} = -\mathbf{L}\mathbf{x} = -\mathbf{L}\mathbf{V}\mathbf{z} = -\mathbf{P}\mathbf{z}$.

Finally, if the plant eigenvalues and those prescribed for the closed-loop system are all distinct from one another, and if the diagonal canonical form is chosen for $\mathbf{A}_{can}$, the Roppenecker's formula appears as a particular case of the present approach. Indeed, the transformation matrix **V** is in this case the modal matrix of the closed-loop system, $\mathbf{V}=\begin{pmatrix}\mathbf{v}_{Li} & \cdots & \mathbf{v}_{Ln}\end{pmatrix}$, which is composed of its eigenvectors, so that (1.79) becomes identical to (1.61) and (1.75) to (1.63). All the remarks of Sect. 1.6 apply thus in this case, in particular those of Sect. 1.6.2 on the targeted choices of the parameter vectors and of the eigenvectors.

Remark 1.17. Differences with the complete modal design.

If for $\mathbf{A}_{can}$ the controllability canonical form is chosen, the necessity, which appears in the complete modal design, to distinguish between the case of all distinct closed-loop eigenvalues and the one of multiple eigenvalues disappears here [BeOs87].

Also, the limitation of the Roppenecker's formula to a multiplicity order of multiple eigenvalues inferior or equal to the number of plant inputs disappears here: any number of identical eigenvalues can be imposed to the closed-loop system if $\mathbf{A}_{can}$ is chosen in controllability canonical form.

1.7.2 Single-input Systems (p = 1): Retrieving Ackermann'sFormula

The parameter matrix **P**, of size (1 × n), becomes then a row vector,

$$\mathbf{P}=\mathbf{p}^{\mathrm{T}}=\begin{pmatrix}p_1 & \cdots & p_n\end{pmatrix}.$$

Starting from (1.76), the initial plant being assumed controllable, which has the consequence in this single-input case that the matrix $\mathbf{Q}_C^{-1}$ exists, we can write that

$$\mathbf{P}\mathbf{D}^{-1}=\mathbf{p}^{\mathrm{T}}\begin{pmatrix}\mathbf{p}^{\mathrm{T}}q_1(\mathbf{A}_{can})\\ \mathbf{p}^{\mathrm{T}}q_2(\mathbf{A}_{can})\\ \vdots\\ \mathbf{p}^{\mathrm{T}}q_n(\mathbf{A}_{can})\end{pmatrix}^{-1}\mathbf{Q}_C^{-1}.$$

Taking into account that

$$\begin{pmatrix}\mathbf{p}^{\mathrm{T}}q_1(\mathbf{A}_{can})\\ \mathbf{p}^{\mathrm{T}}q_2(\mathbf{A}_{can})\\ \vdots\\ \mathbf{p}^{\mathrm{T}}q_n(\mathbf{A}_{can})\end{pmatrix}\begin{pmatrix}\mathbf{p}^{\mathrm{T}}q_1(\mathbf{A}_{can})\\ \mathbf{p}^{\mathrm{T}}q_2(\mathbf{A}_{can})\\ \vdots\\ \mathbf{p}^{\mathrm{T}}q_n(\mathbf{A}_{can})\end{pmatrix}^{-1}=\begin{pmatrix}1 & \cdots & \cdots & 0\\ \vdots & \ddots & & \vdots\\ \vdots & & \ddots & \vdots\\ 0 & \cdots & \cdots & 1\end{pmatrix},$$

and that $q_n(\mathbf{A}_{can}) = 1$ by definition, we deduce from the last line of the above matrix equation that

$$\mathbf{p}^{\mathrm{T}} \begin{pmatrix} \mathbf{p}^{\mathrm{T}} q_1(\mathbf{A}_{can}) \\ \mathbf{p}^{\mathrm{T}} q_2(\mathbf{A}_{can}) \\ \vdots \\ \mathbf{p}^{\mathrm{T}} q_n(\mathbf{A}_{can}) \end{pmatrix}^{-1} = \begin{pmatrix} 0 & \cdots & 0 & 1 \end{pmatrix}.$$

This leads easily to

$$\mathbf{PD}^{-1} = \begin{pmatrix} 0 & \cdots & 0 & 1 \end{pmatrix} \mathbf{Q}_C^{-1} = \mathbf{q}_C^{\mathrm{T}},$$

and the control law (1.75) becomes

$$\mathbf{L} = \mathbf{q}_C^{\mathrm{T}}\, q_0(\mathbf{A}).$$

Since $\mathbf{q}_C^{\mathrm{T}}$ is the last row of the controllability matrix inverse, and since $q_0(\lambda)$ is, according to (1.77), the desired characteristic polynomial of the closed-loop system, we recognize here the Ackermann's formula (1.27).

1.8 State Feedback with Integral Action

It has been seen previously that the static unity gain of a state-feedback control system was ensured by the gain compensation (feedforward) matrix **M**. This solution has of course the inconvenient not to guarantee that the steady-state error vanishes exactly, especially if the plant model is not perfectly known.

An alternative consists in adding one or more integrators to the control loop, exactly as is done for the design of controllers including an integral term (I-term) in the case of SISO systems.

1.8.1 Continuous Case

In order to cancel the steady-state error, let us add q integrators at the output of the comparator, one by reference vector $\mathbf{y}_r$ component, as for the design of cascade controllers in the case of SISO systems, and feed them back by means of a matrix $\mathbf{L}_2$, the state vector **x** of the plant being fed back as previously by a feedback matrix denoted here by $\mathbf{L}_1$.

The control system has then the block diagram of Fig. 1.7:

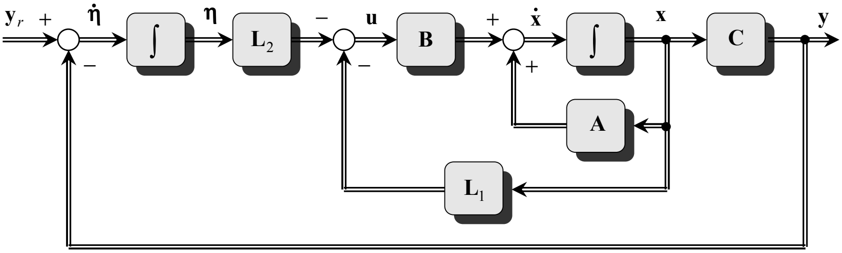


Fig. 1.7 Continuous-time state-feedback control with integral action.

The state vector $\boldsymbol{\eta}$, of size q, reflects the addition of the q integrators (one additional state variable per integrator output). The plant state equations are

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\,\mathbf{u} \\ \mathbf{y} = \mathbf{C}\mathbf{x} \end{cases}$$

As to the additional state variables, they are described by

$$\dot{\boldsymbol{\eta}} = \mathbf{y}_r - \mathbf{y} = \mathbf{y}_r - \mathbf{C}\mathbf{x}\,. \tag{1.84}$$

By grouping these q state variables together with the n state variables describing the plant, we obtain the augmented system described below.

1.8.1.1 Augmented System

It is represented by the differential state equation

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \\ \dot{\boldsymbol{\eta}} = -\mathbf{C}\mathbf{x} + \mathbf{y}_r \end{cases} \tag{1.85}$$

thus by the following augmented state representation:

$$\begin{aligned} \begin{pmatrix} \dot{\mathbf{x}} \\ \dot{\boldsymbol{\eta}} \end{pmatrix} &= \begin{pmatrix} \mathbf{A} & 0 \\ -\mathbf{C} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \boldsymbol{\eta} \end{pmatrix} + \begin{pmatrix} \mathbf{B} \\ 0 \end{pmatrix} \mathbf{u} + \begin{pmatrix} 0 \\ \mathbf{I}_q \end{pmatrix} \mathbf{y}_r \\ \mathbf{y} &= \begin{pmatrix} \mathbf{C} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \boldsymbol{\eta} \end{pmatrix} \end{aligned} \tag{1.86}$$

The state-feedback control law introduced previously (see Fig. 1.7) is of the form

$$\mathbf{u} = -\mathbf{L}_1\,\mathbf{x} - \mathbf{L}_2\,\boldsymbol{\eta} = -\begin{pmatrix}\mathbf{L}_1 & \mathbf{L}_2\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ \boldsymbol{\eta}\end{pmatrix}.$$

The augmented system closed by state feedback obeys then the differential state equation

$$\begin{pmatrix}\dot{\mathbf{x}}\\ \dot{\boldsymbol{\eta}}\end{pmatrix} = \begin{pmatrix}\mathbf{A} & 0\\ -\mathbf{C} & 0\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ \boldsymbol{\eta}\end{pmatrix} - \begin{pmatrix}\mathbf{B}\\ 0\end{pmatrix}\begin{pmatrix}\mathbf{L}_1 & \mathbf{L}_2\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ \boldsymbol{\eta}\end{pmatrix} + \begin{pmatrix}0\\ \mathbf{I}_q\end{pmatrix}\mathbf{y}_r\,,$$

$$\begin{aligned}\begin{pmatrix}\dot{\mathbf{x}}\\ \dot{\boldsymbol{\eta}}\end{pmatrix} &= \begin{pmatrix}\mathbf{A}-\mathbf{B}\mathbf{L}_1 & -\mathbf{B}\mathbf{L}_2\\ -\mathbf{C} & 0\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ \boldsymbol{\eta}\end{pmatrix} + \begin{pmatrix}0\\ \mathbf{I}_q\end{pmatrix}\mathbf{y}_r\\ \mathbf{y} &= \begin{pmatrix}\mathbf{C} & 0\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ \boldsymbol{\eta}\end{pmatrix}\end{aligned} \tag{1.87}$$

1.8.1.2 State Feedback Design

The augmented system (1.86), of size $(n+q)$, being described by

$$\begin{aligned}\begin{pmatrix}\dot{\mathbf{x}}\\ \dot{\boldsymbol{\eta}}\end{pmatrix} &= \widetilde{\mathbf{A}}\begin{pmatrix}\mathbf{x}\\ \boldsymbol{\eta}\end{pmatrix} + \widetilde{\mathbf{B}}\,\mathbf{u} + \begin{pmatrix}0\\ \mathbf{I}_q\end{pmatrix}\mathbf{y}_r\\ \mathbf{y} &= \widetilde{\mathbf{C}}\begin{pmatrix}\mathbf{x}\\ \boldsymbol{\eta}\end{pmatrix},\end{aligned}$$

where

$$\widetilde{\mathbf{A}} = \begin{pmatrix}\mathbf{A} & 0\\ -\mathbf{C} & 0\end{pmatrix},\quad \widetilde{\mathbf{B}} = \begin{pmatrix}\mathbf{B}\\ 0\end{pmatrix}\ \text{and}\ \widetilde{\mathbf{C}} = \begin{pmatrix}\mathbf{C} & 0\end{pmatrix},$$

the present problem consists in determining the *augmented* state-feedback matrix

$$\widetilde{\mathbf{L}} = \begin{pmatrix}\mathbf{L}_1 & \mathbf{L}_2\end{pmatrix} \tag{1.88}$$

so that the closed-loop system, resulting from the control law

$$\mathbf{u} = -\widetilde{\mathbf{L}}\begin{pmatrix}\mathbf{x}\\ \boldsymbol{\eta}\end{pmatrix},$$

and having the system matrix $\widetilde{\mathbf{F}} = \widetilde{\mathbf{A}} - \widetilde{\mathbf{B}}\widetilde{\mathbf{L}}$, has the desired dynamic behavior.

The design can then occur by one of the methods described previously, e.g. by assignment of the $(n+q)$ poles or by any of the other methods presented in Sections 1.4 through 1.7, provided the pair $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ is controllable. [Duc01] has shown that this condition is equivalent to the three following:

1. $\operatorname{rank}\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & 0 \end{pmatrix} = n+q$;
2. the pair $(\mathbf{A}, \mathbf{B})$ is controllable;
3. $p \geq q$, i.e. the number of control inputs must be at least equal to the number of outputs to control.

1.8.1.3 Cancellation of the Steady-state Error with Respect to the Reference at Equilibrium

Let us verify the efficiency of integrator introduction into the control loop by calculating the permanent error of order zero, or static error, at the equilibrium state, $\mathbf{y} - \mathbf{y}_r$, for a constant reference signal $\mathbf{y}_r$. The equilibrium state of the augmented system is characterized by

$$\begin{pmatrix} \dot{\mathbf{x}} \\ \dot{\boldsymbol{\eta}} \end{pmatrix} = 0 . \tag{1.89}$$

The second of these equations, $\dot{\boldsymbol{\eta}} = 0$, leads, according to (1.84) or (1.85), to

$$-\mathbf{C}\mathbf{x} + \mathbf{y}_r = 0 ,$$

thus

$$\mathbf{y} = \mathbf{C}\mathbf{x} = \mathbf{y}_r .$$

1.8.1.4 Elimination of Constant Disturbances at Equilibrium

The addition of integrators in the control loop should also cause the cancellation of a steady-state error stemming from the application of a constant measurement disturbance or a constant load disturbance, since the integrators are placed between the comparator output and the point of application of these disturbances.

Let us verify this property.

The previous diagram is completed, as shown in Fig. 1.8, by a constant measurement disturbance **w**, of size *q*, and by a constant load disturbance **v**, the size of which is unspecified but is taken into account by the matrix **E** so that it can be added to the state equation input.

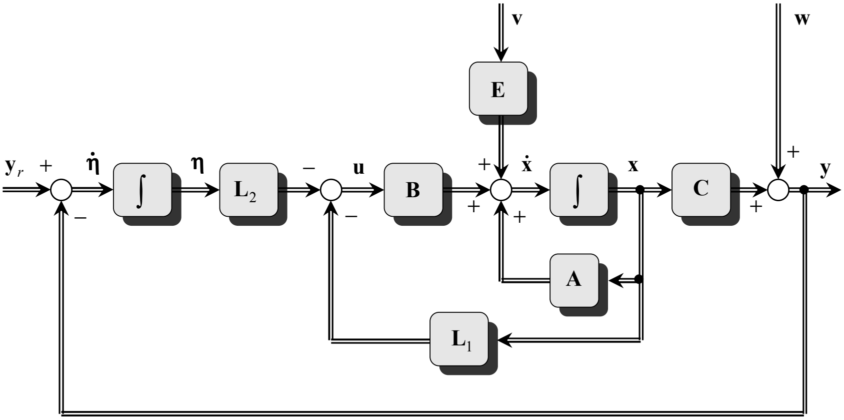


Fig. 1.8 State-feedback control with integral action and load and measurement disturbances.

The augmented system, completed this way, obeys the following state equations:

$$\begin{pmatrix}\dot{\mathbf{x}}\\ \dot{\boldsymbol{\eta}}\end{pmatrix}=\begin{pmatrix}\mathbf{A} & 0\\ -\mathbf{C} & 0\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ \boldsymbol{\eta}\end{pmatrix}+\begin{pmatrix}\mathbf{B}\\ 0\end{pmatrix}\mathbf{u}+\begin{pmatrix}0\\ \mathbf{I}_q\end{pmatrix}\mathbf{y}_r+\begin{pmatrix}\mathbf{E}\\ 0\end{pmatrix}\mathbf{v}-\begin{pmatrix}0\\ \mathbf{I}_q\end{pmatrix}\mathbf{w}$$

$$\mathbf{y}\ =\mathbf{C}\mathbf{x}+\mathbf{w}$$

The closed-loop differential state equation (1.87) is replaced here by

$$\begin{pmatrix}\dot{\mathbf{x}}\\ \dot{\boldsymbol{\eta}}\end{pmatrix}=\begin{pmatrix}\mathbf{A}-\mathbf{B}\mathbf{L}_1 & -\mathbf{B}\mathbf{L}_2\\ -\mathbf{C} & 0\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ \boldsymbol{\eta}\end{pmatrix}+\begin{pmatrix}0\\ \mathbf{I}_q\end{pmatrix}\mathbf{y}_r+\begin{pmatrix}\mathbf{E}\\ 0\end{pmatrix}\mathbf{v}-\begin{pmatrix}0\\ \mathbf{I}_q\end{pmatrix}\mathbf{w}\,.$$

Equilibrium state. The condition (1.89) which characterizes the equilibrium state is equivalent to the two following conditions:

1. $\dot{\mathbf{x}}=0$, which leads to $\mathbf{x}=(\mathbf{A}-\mathbf{B}\mathbf{L}_1)^{-1}\,(\mathbf{B}\mathbf{L}_2\boldsymbol{\eta}-\mathbf{E}\mathbf{v})\neq 0$, thus also to $\mathbf{C}\mathbf{x}\neq 0$
2. $\dot{\boldsymbol{\eta}}=0$, which leads to $-\mathbf{C}\mathbf{x}+\mathbf{y}_r-\mathbf{w}=0$, thus $\mathbf{y}=\mathbf{C}\mathbf{x}+\mathbf{w}=\mathbf{y}_r$.

The constant disturbances are therefore eliminated at equilibrium state.

1.8.2 Discrete Case

The diagram of a discrete system equipped with *q* discrete integrators, added at the comparator output for the same reasons as in the continuous case studied previously, is shown in Fig. 1.9.

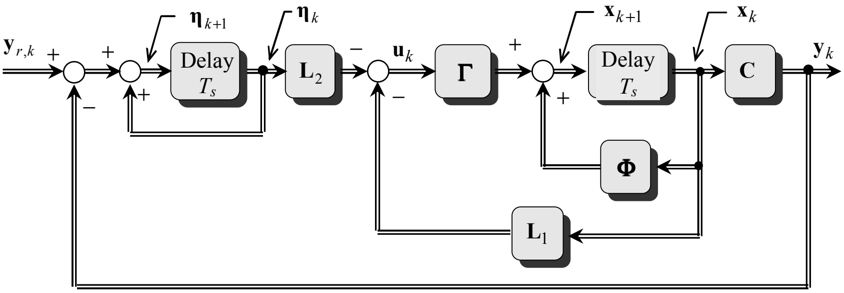


Fig. 1.9 Discrete-time state-feedback control, with *q* discrete integrators.

1.8.2.1 "I" Controller

The addition of a discrete integrator at the comparator output reflects itself in the difference equation

$$\boldsymbol{\eta}_{k+1} = \boldsymbol{\eta}_k + \mathbf{y}_{r,k} - \mathbf{y}_k ,$$

where the integrator output, $\boldsymbol{\eta}$, represents q additional state variables, which add to the n state variables describing the plant.

1.8.2.2 Augmented System State Equations

$$\begin{gathered}\begin{pmatrix}\mathbf{x}_{k+1}\\ \boldsymbol{\eta}_{k+1}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Phi} & 0\\ -\mathbf{C} & \mathbf{I}_q\end{pmatrix}\begin{pmatrix}\mathbf{x}_k\\ \boldsymbol{\eta}_k\end{pmatrix} + \begin{pmatrix}\boldsymbol{\Gamma}\\ 0\end{pmatrix}\mathbf{u}_k + \begin{pmatrix}0\\ \mathbf{I}_q\end{pmatrix}\mathbf{y}_{r,k} = \widetilde{\boldsymbol{\Phi}}\begin{pmatrix}\mathbf{x}_k\\ \boldsymbol{\eta}_k\end{pmatrix} + \widetilde{\boldsymbol{\Gamma}}\mathbf{u}_k + \begin{pmatrix}0\\ \mathbf{I}_q\end{pmatrix}\mathbf{y}_{r,k}\\ \mathbf{y}_k = \begin{pmatrix}\mathbf{C} & 0\end{pmatrix}\begin{pmatrix}\mathbf{x}_k\\ \boldsymbol{\eta}_k\end{pmatrix} = \widetilde{\mathbf{C}}\begin{pmatrix}\mathbf{x}_k\\ \boldsymbol{\eta}_k\end{pmatrix}\end{gathered} \tag{1.90}$$

1.8.2.3 State Feedback Control

$$\mathbf{u}_k = -\mathbf{L}_1\mathbf{x}_k - \mathbf{L}_2\boldsymbol{\eta}_k = -\begin{pmatrix}\mathbf{L}_1 & \mathbf{L}_2\end{pmatrix}\begin{pmatrix}\mathbf{x}_k\\ \boldsymbol{\eta}_k\end{pmatrix},$$

thus

$$\begin{pmatrix}\mathbf{x}_{k+1}\\ \boldsymbol{\eta}_{k+1}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Phi}-\boldsymbol{\Gamma}\mathbf{L}_1 & -\boldsymbol{\Gamma}\mathbf{L}_2\\ -\mathbf{C} & \mathbf{I}_q\end{pmatrix}\begin{pmatrix}\mathbf{x}_k\\ \boldsymbol{\eta}_k\end{pmatrix} + \begin{pmatrix}0\\ \mathbf{I}_q\end{pmatrix}\mathbf{y}_{r,k} \tag{1.91}$$

1.8.2.4 Static Error at Equilibrium

The equilibrium state is characterized here by $\boldsymbol{\eta}_{k+1} = \boldsymbol{\eta}_k$, thus, according to the second equation of (1.91), $0 = -\mathbf{C}\mathbf{x}_k + \mathbf{y}_{r,k}$, i.e. $\mathbf{y}_k = \mathbf{y}_{r,k}$.

1.8.3 Partial Integral Action (Continuous or Discrete)

In the two previous sections, it was assumed that the output vector was fed back entirely, by means of one or more integrators introduced in the control loop.

It may happen that this "I-type" control is wished for only part of the measured quantities, i.e. those for which exact cancellation of the steady-state error is most critical. This limitation can stem from, e.g., the wish not to increase too much the number of state variables, in the case of a large size MIMO system, or not to penalize too much its transient response by the addition of the corresponding poles.

The choice of the q_i components of the output vector $\mathbf{y}(t)$ or $\mathbf{y}_k$ which one wants to feed back by integrators is made by a selection matrix $\mathbf{S}_i$, as illustrated in the two universal simulation diagrams of Appendix D, so that the partial output vector which is fed back is given by

$$\mathbf{y}_i = \mathbf{S}_i\mathbf{y} = \mathbf{S}_i\mathbf{C}_m\mathbf{x} = \mathbf{C}_i\mathbf{x} \ ,$$

with $\mathbf{C}_i = \mathbf{S}_i\mathbf{C}_m$. Even though the method applies equally to the discrete case, the following calculations will be developed only in the continuous case. If $\boldsymbol{\eta}$, of size q_i, denotes as previously the vector of added state variables, the open-loop augmented-system state equation (1.86) becomes

$$\begin{pmatrix} \dot{\mathbf{x}} \\ \dot{\boldsymbol{\eta}} \end{pmatrix} = \begin{pmatrix} \mathbf{A} & 0 \\ -\mathbf{C}_i & 0 \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \boldsymbol{\eta} \end{pmatrix} + \begin{pmatrix} \mathbf{B} \\ 0 \end{pmatrix} \mathbf{u} + \begin{pmatrix} 0 \\ \mathbf{S}_i \end{pmatrix} \mathbf{y}_r \ ,$$

its output equation remaining unchanged.

The application to this system of the control law illustrated in the simulation diagram and including the gain compensation resulting from the matrix **M,** or $\mathbf{S}_c\mathbf{M}$ if $p < q$ as discussed in Sect. 1.1.3.2,

$$\mathbf{u} = -\begin{pmatrix} \mathbf{L}_1 & \mathbf{L}_2 \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \boldsymbol{\eta} \end{pmatrix} + \mathbf{M}\mathbf{y}_r \ ,$$

yields for the closed-loop system the following state equation:

$$\begin{pmatrix}\dot{\mathbf{x}}\\ \dot{\boldsymbol{\eta}}\end{pmatrix}=\begin{pmatrix}\mathbf{A}-\mathbf{BL}_1 & -\mathbf{BL}_2\\ -\mathbf{C}_i & 0\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ \boldsymbol{\eta}\end{pmatrix}+\begin{pmatrix}\mathbf{BM}\\ \mathbf{S}_i\end{pmatrix}\mathbf{y}_r\,.$$

Determination and role of the M matrix. In steady state, the response to a constant reference $\mathbf{y}_r$ gives then

$$0=(\mathbf{A}-\mathbf{BL}_1)\mathbf{x}_\infty-\mathbf{BL}_2\boldsymbol{\eta}_\infty+\mathbf{BMy}_r\,,$$

where $\mathbf{x}_\infty$ and $\boldsymbol{\eta}_\infty$ denote the state and its augmented part at equilibrium. Accordingly, the output at equilibrium has the following expression:

$$\mathbf{y}_\infty=\mathbf{Cx}_\infty=-\mathbf{C}(\mathbf{BL}_1-\mathbf{A})^{-1}\mathbf{BL}_2\boldsymbol{\eta}_\infty+\mathbf{C}(\mathbf{BL}_1-\mathbf{A})^{-1}\mathbf{BMy}_r\,.$$

Let us write, similarly to (1.8), $\mathbf{M}=\rho\left[\mathbf{C}(\mathbf{BL}_1-\mathbf{A})^{-1}\mathbf{B}\right]^{-1}$, with $0\le\rho\le 1$, respectively $\mathbf{M}=\rho\,\mathbf{S}_c\left[\mathbf{C}(\mathbf{BL}_1-\mathbf{A})^{-1}\mathbf{B}\right]^{-1}$ if $p<q$.

Due to the presence of the integrators, $\boldsymbol{\eta}_\infty$ is adjusted automatically so that $\mathbf{S}_i\,\mathbf{y}_\infty=\mathbf{S}_i\,\mathbf{y}_r$. In the case where $\rho=1$, situation which has been chosen in the accompanying software *MMCE.m*, the equality $\mathbf{y}_\infty=\mathbf{y}_r$, which leads to $\mathbf{S}_i\,\mathbf{y}_\infty=\mathbf{S}_i\,\mathbf{y}_r$, is obtained for $\boldsymbol{\eta}_\infty=0$. The anticipation term, or feedforward term, constituted by the **M** matrix yields then at equilibrium the maximum contribution to the required control, the integrators having no effect anymore, in the hypothesis of a perfect model and in the absence of any disturbance. Due to its immediate application after a step reference, this term accelerates the transient response.

In counterpart, it causes generally also overshoots of this response, due to the zero(s) that it introduces into the closed loop, exactly as does a feedforward compensation term in SISO control loops.

Its influence can be reduced by choosing $\rho<1$, but in this case the equality $\mathbf{y}_\infty=\mathbf{y}_r$ can no longer be obtained, unless $\mathbf{C}_i=\mathbf{C}$, i.e. $\mathbf{S}_i=\mathbf{I}_q$.

An application is proposed in the frame of the solved exercise dealing with the control of a magnetic tape drive, at the end of Chapter 4.

1.9 Solved Exercises

Preliminaries and Computer-based Calculations. Most of the exercises proposed in this book require the use of digital tools to be solved. Due to the importance and universal acceptance of the MATLAB® [2] software environment in this field, a generic program, created under MATLAB 7.9 (R2009b) but compatible with earlier versions and covering the five plants which will be used in the exercises, is freely available to the reader. It contains a main program *MMCE.m* and several subprograms, bound in the archive *mmce.zip*, which is downloadable on Internet from the Springer or The MathWorks sites:

http://extras.springer.com/2010/978-3-642-13733-4
http://www.mathworks.com/matlabcentral/fileexchange.

It enables the reader, eager to solve the exercises by himself, to compare his solutions with the given ones, and even to find other solutions, knowing the multiplicity of choices offered in MIMO control. The limited place available in this book has indeed not made it possible to apply them all on each of the included plants.

Simulation makes this learning process even more vivid, and the reader is never advised enough to undertake simulations of his solutions. Three Simulink® simulation diagrams, which enable repeating the simulations of about all the exercises, are given in Appendix D. They can also be used as templates to create other diagrams, at the reader's liking.

Despite all the care brought to establishing software modules or diagrams, no warranty can be given as to the absence of any error. It is therefore called upon the reader's perspicacity in his use of these tools.

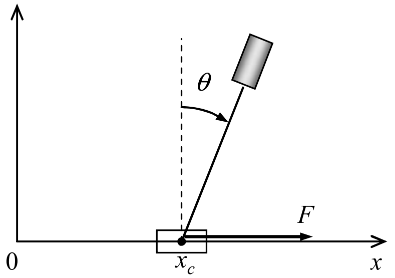
Exercise 1.1 Control of an Inverted Pendulum

The plant to study is sketched in Fig. 1.10. It represents a real experimental setup, which will be used repeatedly along this book [Ami92]. Thus the real parameters of this setup will be used in the present exercise as well as in the followings, and their values appear in the generic script *m*-file *MMCE.m*, under the plant name Inverted pendulum LIP 100 (Amira).

The pendulum rotates about an axis, carried by a cart, which is moved horizontally by a belt and a motor, not shown in the figure, which apply a force *F* to it.

The mathematical representation of such a system has been established in several books, typically: in [MoMa98] by the method of the Lagrangian equations, described e.g. in [Ost04], by neglecting all the frictions; in [FrPE94] where the viscous friction of the cart on its rail is taken into account by writing the equilibrium of forces and the fundamental equation of dynamics; in [Ami92] where the viscous friction of both the translation movement and the rotation movement are taken into account, also by application of the fundamental principle of dynamics and the balance of forces involved. It is this last situation which will be considered here.

[2] Distributed by The MathWorks. The only toolbox required to work out the exercises is the *Control System Toolbox*. To do the proposed simulations, it is necessary to have also *Simulink*®.



x_c : cart position
θ: pendulum angle
F : force applied to the cart

Fig. 1.10 Inverted pendulum on a cart.

The model equations, linearized around its unstable equilibrium point in the small angle approximation, are the following[3]:

$$\begin{cases}(M_c+M_p)\ddot{x}_c+M_p\ell\ddot{\theta}+B_c\dot{x}_c=F\\(J+M_p\ell^2)\ddot{\theta}+M_p\ell\ddot{x}_c+B_p\dot{\theta}-M_p\ell g\theta=0\end{cases}$$

where the parameters have the following meanings:

$M_c, M_p =$ masses of the cart and the cylinder mounted at the rod extremity;

$B_c, B_p =$ viscous friction coefficients of the linear cart movement and the rotational rod movement;

$\ell, J =$ length and moment of inertia of the rod.

Choosing the following state variables and input quantity:

$$x_1=x_c;\ x_2=\theta;\ x_3=\dot{x}_c;\ x_4=\dot{\theta};\ u=F\ ,$$

and expressing all physical quantities in terms of their electric tensions, delivered by the sensors or applied to the electronic circuit driving the motor, the state model of the plant [Ami92] in continuous time is given by

$$\mathbf{A}=\begin{pmatrix}0 & 0 & -1.950 & 0\\0 & 0 & 0 & 1\\0 & -0.1289 & -1.915 & 0.00082\\0 & 21.47 & 26.34 & -0.1362\end{pmatrix};\quad \mathbf{b}=\begin{pmatrix}0\\0\\-6.1344\\84.303\end{pmatrix}.$$

This plant is sampled at a period $T_s=0.03$ s, and the discrete model is given by

[3] In certain books it is the positive trigonometric direction which is used as positive angle direction, and not the clockwise direction as is the case here. It is enough then to inverse the sign of all terms in θ , $\dot{\theta}$ or $\ddot{\theta}$.

$$\mathbf{\Phi} = e^{\mathbf{A}T_s} = \begin{pmatrix} 1 & 1.11\times10^{-4} & -5.68\times10^{-2} & 4.1\times10^{-7} \\ 0 & 1.01 & 1.16\times10^{-2} & 3.0035\times10^{-2} \\ 0 & -3.762\times10^{-3} & 0.944 & -3.303\times10^{-5} \\ 0 & 0.6435 & 0.769 & 1.0056 \end{pmatrix};$$

$$\mathbf{\Gamma} = \int_0^{T_s} e^{\mathbf{A}\tau} d\tau \cdot \mathbf{b} = \begin{pmatrix} 0.0053 \\ 0.0372 \\ -0.1789 \\ 2.461 \end{pmatrix}.$$

The state variables x_1 and x_2 will be selected here as output quantities, although x_3 (but not x_4) is also measurable. The *measurement* matrix is therefore

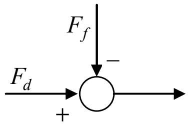
$$\mathbf{C}_m = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

The notation $\mathbf{C}_m$ will be used in all this book's exercises solved by *MMCE.m* for the matrix building the real measurements. An identity matrix **C** is indeed introduced in the simulation diagrams inside the blocks representing the plants in state-space representation, so as to give full access to their state vector.

The reader is free to modify this measurement matrix $\mathbf{C}_m$ at will, in order to perform other tests.

Taking into account nonlinear phenomena. The most significant nonlinearities are the dry friction force (Coulomb friction, of intensity F_c) to which the cart is submitted during its movement on the rail, and the stiction force (static friction, of intensity F_s) which appears at zero speed. The total friction force F_f which counteracts the driving force, F_d, as sketched below, is thus given by:

$$F_f = \begin{cases} \operatorname{sign}(F_d)\cdot\min(|F_d|, F_s), & \text{for } \dot{x} = 0, \\ F_c \operatorname{sign}[\dot{x}(t)], & \text{for } \dot{x} \neq 0\,. \end{cases}$$



In this case a compensation can be applied by superimposing an additional input tension which amounts to $u_{s0} = F_f / K_F$, where K_F is the conversion factor expressed in N/V. The quantity u_{s0} can be either evaluated once and for all by measurements on the real set-up, or estimated in real time by means of a disturbance observer. This second method will be the subject of an exercise in Chap. 2.

In the present exercise, we will not take into account this nonlinearity and will consider the plant as a linear system corresponding to the above state space model.

This exercise will be solved in the MATLAB® environment, with the *MMCE.m* program mentioned above.

a) Determine the pole-zero map of the open loop.

b) It is desired to stabilize the pendulum in its upright position, the cart being stopped at a reference position $y = x_1 = y_r$ on its rail. After having checked the controllability of the system, design a state-feedback control for it by the pole placement method, by assuming first that the state vector is fully accessible and by imposing to the closed-loop system the following multiple eigenvalue: $\lambda_{Li} = -5\,\mathrm{s}^{-1}, \quad i = 1, \cdots, 4$.
Determine the poles and zeros of the closed-loop system and discuss.

c) Simulate this control system and log its step response. Observe in particular the steady-state values.

d) Switch to the discrete model and repeat questions b) and c).

Solution:

Run the generic program *MMCE.m* and select the plant inverted pendulum LIP100 (Amira) with two measurements.

(a) The statement [Poles Zeros] = pzmap(A,B,Cm(1,:),D(1,:)) included in the software yields the following values for the open-loop plant, from its unique input u to its output $y = x_1 = x_c$:
poles: $s_1 = 0,\ s_2 = 4.51,\ s_3 = -4.83,\ s_4 = -1.73$; zeros: $z_1 = -4.50,\ z_2 = 4.38$.

(b) In the program, select then successively: Choice of model: Continuous (default), Synthesis of state-feedback (default), Integral action: 0 = without (default), Component of y to be set to its reference value: 1, then Algorithm to be used: 1) pole placement (MATLAB modules « acker.m » or « place.m »).

Before the last prompt of the program, the rank of the controllability matrix $\mathbf{Q}_C = \left(\mathbf{b} \quad \mathbf{Ab} \quad \mathbf{A}^2\mathbf{b} \quad \mathbf{A}^3\mathbf{b}\right)$ has been computed by the statement rank(ctrb(A,B)): it is equal to 4, thus to the order n of the plant, which is therefore controllable.

Since the plant has less control inputs than measurement outputs ($p < q$), the program prompts the user to select the output or outputs which should be regulated without steady-state error to a constant input reference. Answer here: 1, so as to obtain a unit static gain on the *position* output ($y = y_1 = x_1$).

Since we have entered a multiple eigenvalue with a multiplicity order greater than the number of control inputs, the algorithm *acker.m* (Ackermann's method) is chosen automatically by the program. The obtained result is:

$$\mathbf{L} = \boldsymbol{\ell}^{\mathrm{T}} = \begin{pmatrix} -2.652 & 2.449 & 4.485 & 0.539 \end{pmatrix}$$
$$M = -2.652$$

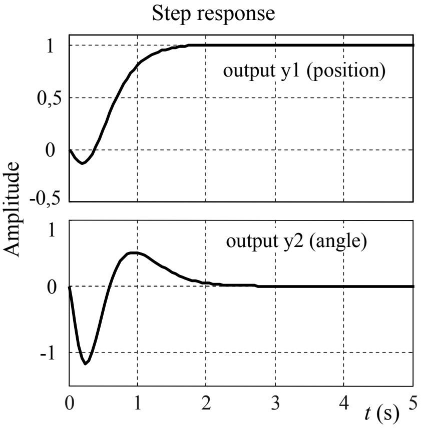
The closed-loop poles and transmission zeros from y_r to $y = x_1$ are found by executing the statement pzmap(sys_CL), the closed-loop system being built by sys_CL = ss(A_CL, B_CL, Cm(1:p,:),D(1:p,:)), with $\mathbf{A}_{CL}$ and $\mathbf{B}_{CL}$ given by (1.4) and (1.5):

$$\text{poles: } s_1 = -5.0008,\ s_{2,3} = -5.0000 \pm j0.0008,\ s_4 = -4.9992;$$
$$\text{zeros: } z_1 = -4.50,\ z_2 = 4.38\,.$$

The poles are at the prescribed locations, but a small inaccuracy is noticeable, which is characteristic of Ackermann's method, in particular in the case of high order plants and multiple poles, which is the case of the present selection. The *place.m* algorithm also proposed by the program is more robust, but it does not allow entering multiple eigenvalues with a multiplicity order greater than the size of the input signal (equal to 1 here, the control u being scalar).

The zeros remain unchanged as we have seen in Sect. 1.2.4. The existence of a zero in the right-half complex plane, often also called improperly *unstable zero* though it does not have any effect on the system stability, reveals on the contrary the *nonminimal phase* characteristic of the plant, as well in open loop as in closed loop.

(c) The closed-loop step response is plotted by using the statements at the end of the program and is reproduced in the oscillogram to the right. The reader having the Simulink® software at his disposal can retrieve this response also by means of the generic simulation diagram given in Appendix D. The $\mathbf{S}_c$ matrix of this diagram extracts from the reference $\mathbf{y}_r = \begin{bmatrix} 1 & 0 \end{bmatrix}$ the unique component $y_{r,1}$ for which the M matrix (a scalar, here) can ensure a unit static gain (see the third paragraph of the answer to question (b) above and the discussion of Sect. 1.1.3.2, *Case where* $p \neq q$).



In the transient part of the output y_1 step response we observe an undershoot, which is characteristic of nonminimal phase systems. The steady-state response corresponds to a unit static gain, which is obtained by the gain compensation, or feedforward, matrix *M*, which is here a scalar.

(d) Select now the option Choice of model: discrete, and proceed as previously. The closed-loop eigenvalues to enter are the *continuous-time values*, the program converting them automatically to their discrete counterpart at the plant sampling time, here $T_s = 0.03$ s. The obtained discrete control law is

$$\ell^{\mathrm{T}} = \begin{pmatrix} -2.0298 & 2.0162 & 3.5714 & 0.4434 \end{pmatrix}; \quad M = -2.0298\,.$$

Closed-loop poles and zeros: $p_i = 0.8607 = e^{-0.03\times 5}, \quad i = 1,\ldots,4$

$$z_1 = -0.9809\,,\ z_2 = 0.8737\,,\ z_3 = 1.1403$$

We note the presence of an additional zero as compared with the continuous case: it results from the sampling of the continuous plant *with zero-order hold.*

The step responses are identical to the previous ones, at the sampling times.

Exercise 1.2 Control of a Three-tank System

The considered MIMO plant is sketched in Fig. 1.11[4].The state variables x_1, x_2 and x_3, respectively the liquid feeding flows u_1 and u_2 represent the deviations of the three liquid levels from their reference heights, respectively from their reference flow rates, which correspond themselves to a given equilibrium state.

If we assume that only the level deviations x_1 and x_3 are measured, the state representation of this plant is

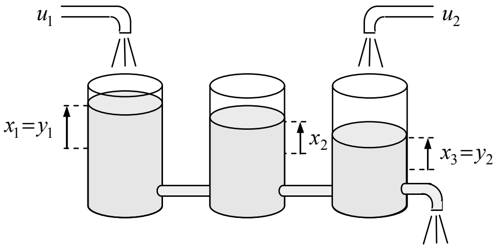


Fig. 1.11 Three-tank system.

[4] Numerical example introduced in [Föl90] and reused in [Rop90].

$$\mathbf{A}=\begin{pmatrix}-0.332 & 0.332 & 0\\ 0.332 & -0.664 & 0.332\\ 0 & 0.332 & -0.524\end{pmatrix},\quad \mathbf{B}=\begin{pmatrix}0.764 & 0\\ 0 & 0\\ 0 & 0.764\end{pmatrix},$$

$$\mathbf{C}_m=\begin{pmatrix}1 & 0 & 0\\ 0 & 0 & 1\end{pmatrix},$$

the liquid heights being expressed in meters and the flow rates in m^3/mn. The symmetry of matrices **A** and **B** is a consequence of the equality of the tank and pipe cross areas.

a) Calculate by the simple modal control method available in the program a continuous-time state feedback which shifts the two eigenvalues corresponding to the two slowest modes to the following values: $-0.5\,\mathrm{mn}^{-1}$; $-2\,\mathrm{mn}^{-1}$.

Plot or observe by simulation the free response of the three state variables, starting from an initial state $x_0=\begin{pmatrix}-0.8 & -0.5 & -0.2\end{pmatrix}^{\mathrm{T}}$, for the open-loop plant, then for the system closed by the controller calculated in (a).

b) Plot or observe by simulation the four individual step responses from input *i* to output *j*, for $i, j = 1, 2$. In order to better separate the steady-state responses, steps of different amplitudes, different from unity, will be applied by the program to the two inputs: $y_r = 0.4$ and $y_r = 0.8$. What do you observe?

c) Repeat this design by using now the Falb-Wolovich input-output decoupling algorithm, and compare the resulting step responses with those of question (b). If possible, use the same eigenvalues as imposed in (a).

d) Determine the zeros of the closed-loop system, and compare them with its eigenvalues.

Solution:

Run the generic program *MMCE.m* by choosing the three-tank system plant.

(a) The program calculates the initial plant eigenvalues:

$$\lambda_1=-1.0359,\ \lambda_2=-0.4374,\ \lambda_3=-0.0467.$$

Since the plant has two control inputs, only two eigenvalues can be given new values by this method. It is then usual to choose the slowest ones, so we decide to place λ_2 at $\lambda_{L2} = -2$ and λ_3 at $\lambda_{L3} = -0.5$. The state feedback matrix which achieves this is then

$$\mathbf{L} = \begin{pmatrix} 1.0717 & 0.1280 & -0.5820 \\ -0.8003 & 0.6388 & 1.5670 \end{pmatrix}.$$

The free response, or return to equilibrium at vanishing input from a non vanishing initial state, $x_0 = \begin{pmatrix} -0.8 & -0.5 & -0.2 \end{pmatrix}^{\mathrm{T}}$, is plotted in Fig. 1.12, for the initial plant and the closed loop. The two time scales have been chosen different so as to enlighten better the time reduction achieved by the regulation in eliminating the initial disturbance.

A verification by simulation shows that the two control signals applied to the plant inputs during the transient response do not exceed $1\ \mathrm{m}^3\ /\ \mathrm{mn}$, which represents the maximum supposed flow rate of the actuators in this example. If this were not the case, we should choose closed-loop eigenvalues closer to the imaginary axis, which would of course result in a slower transient response.

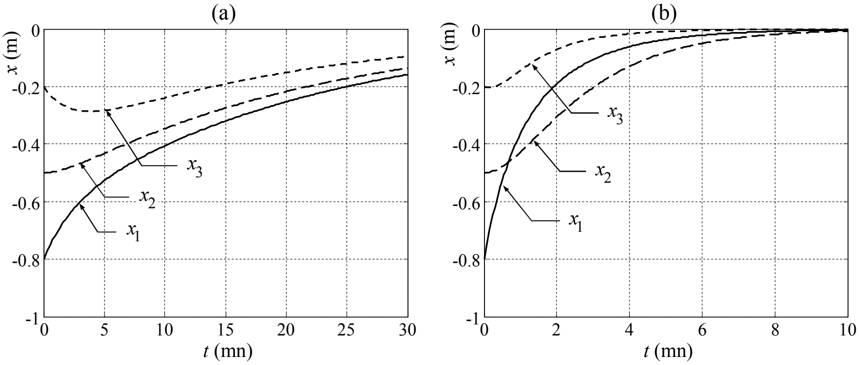


Fig. 1.12 Free responses of the state variables: (a) open-loop; (b) closed-loop.

(b) The step responses to the references given in the problem statement are plotted in Fig. 1.13.

Thanks to the feedforward matrix **M** which has been calculated by the program at question (a),

$$\mathbf{M} = \begin{pmatrix} 1.3529 & -0.7353 \\ -0.6981 & 2.3550 \end{pmatrix},$$

a steady-state unit gain is ensured between each of the references and the corresponding output. As to the transient behavior, the signal applied to input 1 has a visible influence on output 2 and vice-versa. This reveals a coupling inside the plant.

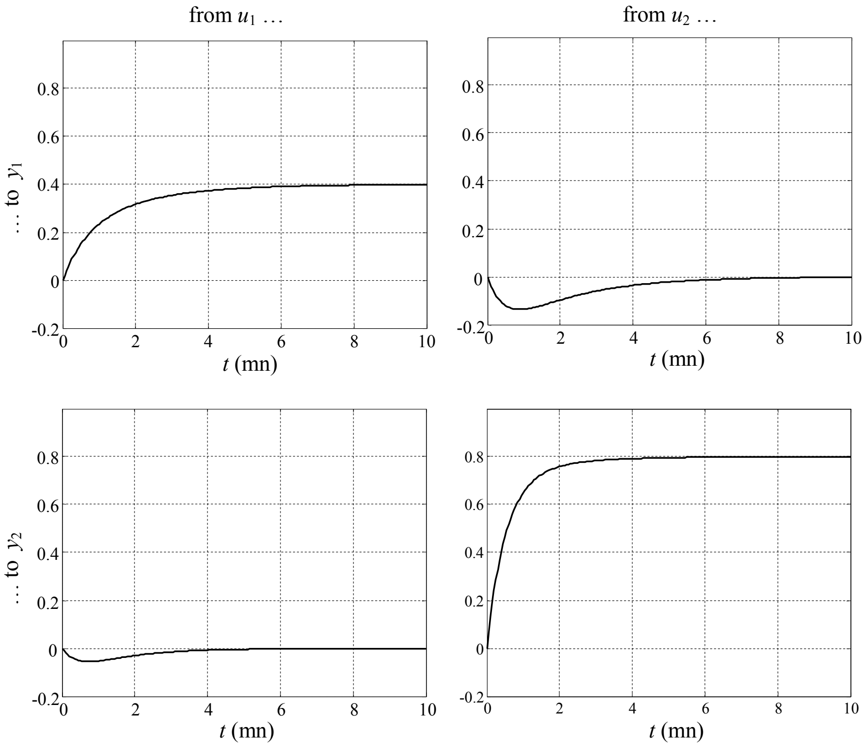


Fig. 1.13 Step responses of the closed-loop system.

(c) The design occurs here by the first of the two methods available in the program under decoupling method (Falb-Wolovich or Roppenecker).

The order differences of the two outputs, δ_1 and δ_2, are both equal to 1, since

$$\mathbf{c}_1^{\mathrm{T}}\mathbf{B}=\begin{pmatrix}1 & 0 & 0\end{pmatrix}\mathbf{B}=\begin{pmatrix}0.764 & 0\end{pmatrix}\neq 0$$
$$\mathbf{c}_2^{\mathrm{T}}\mathbf{B}=\begin{pmatrix}0 & 0 & 1\end{pmatrix}\mathbf{B}=\begin{pmatrix}0 & 0.764\end{pmatrix}\neq 0\,.$$

The total order difference of the plant amounts therefore to $\delta=\delta_1+\delta_2=2$, and only 2 eigenvalues can be imposed to the closed loop. Let us use the same eigenvalues as in question (a), for comparison purposes: -2 and -0.5. The following control law and feedforward matrices are obtained here:

$$\mathbf{L}=\begin{pmatrix}2.1832 & 0.4346 & 0\\ 0 & 0.4346 & -0.0314\end{pmatrix};\quad \mathbf{M}=\begin{pmatrix}2.6178 & 0\\ 0 & 0.6545\end{pmatrix}. \qquad (1.92)$$

The step responses are given in Fig. 1.14. One will notice that the couplings have disappeared, each of the reference signals acting only upon its associated output.

The closed-loop system is decoupled, as confirmed by its transfer matrix $\mathbf{G}(s)$, calculated by the statement Gw = tf(sys_CL) of the program:

$$\mathbf{G}(s) = \begin{pmatrix} \dfrac{2}{s+2} & 0 \\ 0 & \dfrac{0.5}{s+0.5} \end{pmatrix}.$$

Among its poles are as expected the two eigenvalues imposed during the design.

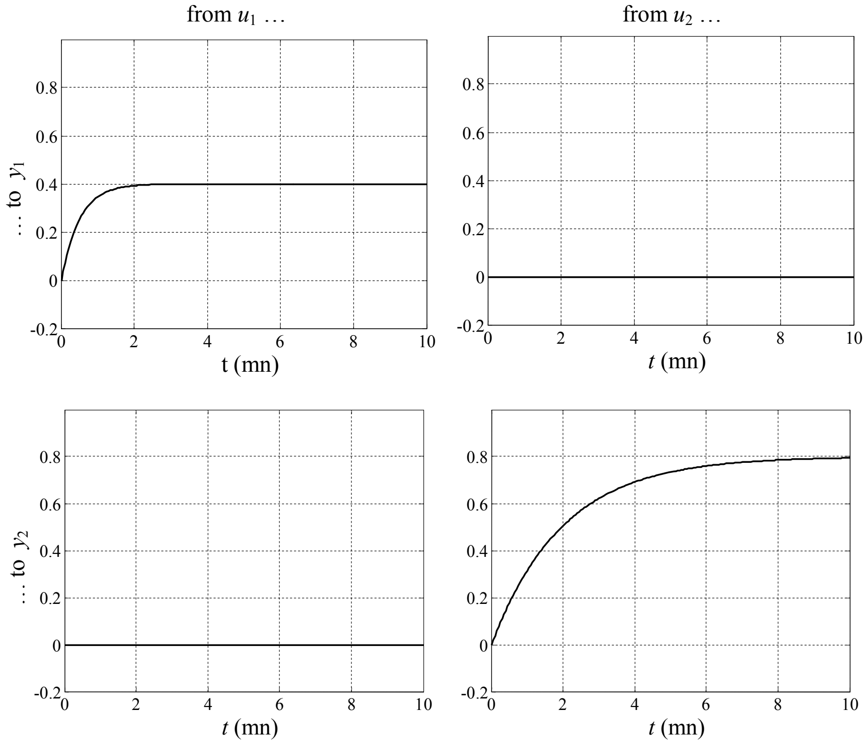


Fig. 1.14 Closed-loop step responses with decoupling control.

(d) The closed-loop system has one zero, given by Z = zero(sys_CL): $z_1 = -0.664$.

Its eigenvalues, given by eig(sys_CL), are: -2, -0.5 and -0.664. It appears thus that the third closed-loop eigenvalue has been placed by this design exactly at the location of the zero of the system, open or closed-loop since it is the same, which has had the effect to cancel out this pole from the transfer matrix of the decoupled system.

This is confirmed by using the statement Msys_CL = minreal(sys_CL) of the program, which answers by indicating that one state has been removed. This statement builds indeed the minimal representation of a system by suppressing all uncontrollable and/or unobservable modes (see Sections A.2.3 and A.3.3 of Appendix A).

Exercise 1.3 Complete Modal Control of a Three-tank System

Consider again the three-tank system of the previous exercise.

a) Take advantage of the degrees of freedom offered by the complete modal control to realize a state-feedback control which decouples the inputs-outputs. Comment the differences with the Falb-Wolovich method.

b) The middle tank has now a hole through which a constant liquid flow leaks, with a flow rate of 0.5 m^3/mn. By considering this flow as a constant disturbance *v*, check if it is possible to calculate a state-feedback control which decouples this disturbance from the outputs, and, if the answer is positive, observe by simulation the effect of the disturbance on the closed-loop system outputs.

Solution:

Run the *MMCE.m* program by selecting the synthesis of a state feedback by the decoupling method and choosing this time its second variant (method of Roppenecker).

(a) As has been seen in Sect. 1.6.5, the choice of the eigenvalues and the parameter vectors is not arbitrary, if one wants a decoupling of the closed-loop inputs and outputs. These choices are going to be applied in this module. The first step consists in looking for eventual zero or zeros of the initial plant. The statement [Poles Zeros] = pzmap(Ac, Bc, Cmm(1:p,:), D(1:p,:)) of the main program has given a zero at $\mu = -0.664$. One eigenvalue must then necessarily be placed at $\lambda_{L1} = -0.664$: this is done automatically by the subprogram *Falb_Wolovich_or_Roppenecker.m*. One chooses then freely $\lambda_{L2} = -2$ and $\lambda_{L3} = -0.5$ so as to provide for comparison of the results with those of Exercise 1.2.

The program determines then the directions associated with this zero, i.e. the non vanishing solutions of the homogeneous equation constructed with the Rosenbrock matrix (Appendix A, Sect. A.6):

$$\begin{pmatrix} \mathbf{A}-\mu\mathbf{I} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \begin{pmatrix} \mathbf{x}_Z \\ \mathbf{u}_Z \end{pmatrix} = \mathbf{0}\,.$$

The five statements Rosen = [A-Z_OL(i)*eye(n),B;C,D], nsp = null(Rosen,'r'), nv = size(nsp,2), Xz = nsp(1:n,1:nv), and Uz = nsp(n+1:n+p,1:nv) yield

$$\mathbf{x}_Z = \begin{pmatrix} 0 & -2.3012 & 0 \end{pmatrix}^{\mathrm{T}} \text{ and } \mathbf{u}_Z = \begin{pmatrix} 1 & 1 \end{pmatrix}^{\mathrm{T}},$$

which determine the eigenvector and the parameter vector associated with the eigenvalue $\lambda_{L1} = \mu$, according to Step 1 of the procedure of Sect. 1.6.5: $\mathbf{v}_{L1} = \mathbf{x}_Z$ and $\mathbf{p}_1 = -\mathbf{u}_Z$. The two other parameter vectors are calculated according to (1.71)

$$\mathbf{p}_1 = \begin{pmatrix} 2.0753 & -0.10799 \end{pmatrix}^{\mathrm{T}},\ \mathbf{p}_2 = \begin{pmatrix} 0.8797 & 0.8483 \end{pmatrix}^{\mathrm{T}}.$$

The final result is exactly the same control law (1.92) as was obtained by the method of Falb-Wolovich used in Exercise 1.2, question (c).

There is thus no difference between these two designs, in the present case. If, on the contrary, the plant zero had been situated in the right-half plane, the algorithm 3) complete modal synthesis (Roppenecker's formula) of the program would have given us the choice not to compensate it by an imposed eigenvalue, and to perform only a partial decoupling, what the Falb-Wolovich method does not allow.

(b) The second tank leak represents a constant load disturbance v applied to the plant by means of the vector $\mathbf{e} = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}^{\mathrm{T}}$. The feasibility necessary condition (1.73) of such a synthesis, namely

$$\mathbf{C}_m\,\mathbf{e} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}^{\mathrm{T}} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

is satisfied. Furthermore, the plant has indeed a zero and the input vector $\mathbf{e}$ can be written as a multiple of the state direction associated with this zero, $\mathbf{x}_Z = \begin{pmatrix} 0 & -2.3012 & 0 \end{pmatrix}^{\mathrm{T}}$. The envisioned synthesis is thus possible. One eigenvalue must be placed at the location of the zero and the corresponding parameter vector must be chosen equal to the direction of the control vector associated with this zero. The two other pairs of eigenvalues and parameter vectors can be chosen according to other criteria.

Since the design of the $\mathbf{L}$ matrix performed at question (a) fulfills precisely these various requirements, the control law obtained in that question can be used here. The oscillograms of Fig. 1.15 represent the step responses, with the application of the disturbance having a 0.5 m^3/mn amplitude at time $t = 15$ mn. By comparison, the same responses are plotted for a complete modal synthesis (Roppenecker's formula), performed with the same eigenvalue choice as in question (a), including the mandatory choice of $\lambda_{L1} = -0.664$ which is not done automatically here by the program, but with randomly chosen parameter vectors: $\begin{pmatrix}1 & 2\end{pmatrix}^{\mathrm{T}}$, $\begin{pmatrix}0 & 1\end{pmatrix}^{\mathrm{T}}$ and $\begin{pmatrix}1 & 0\end{pmatrix}^{\mathrm{T}}$.

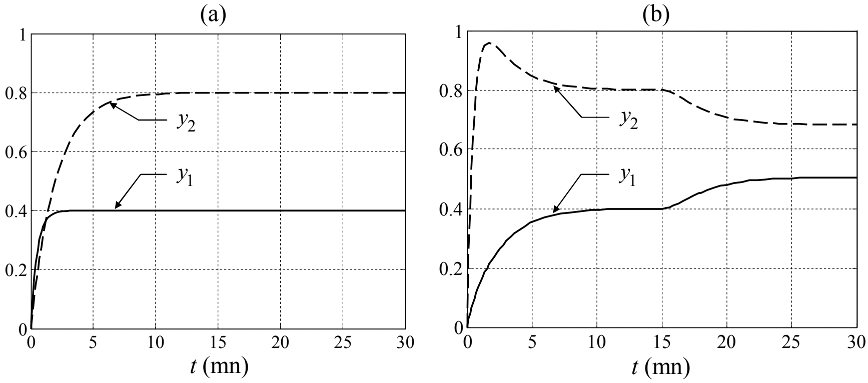


Fig. 1.15 Simulation oscillograms obtained with the
(a) decoupling method according to Roppenecker;
(b) complete modal synthesis (Roppenecker's formula), with random parameter vectors.

It appears clearly that in case (b), the disturbance-output decoupling is lost.

Exercise 1.4 State Feedback with Integral Action

The three-tank set-up of the two previous exercises is considered again, but the study will be made here in discrete time, with a sampling period of 0.1 minute.

a) Calculate for this plant a control law with integral action, which gives the closed-loop system the same continuous-time eigenvalues as in Exercise 1.2, plus additional eigenvalues to choose judiciously. Choose freely one of the available design methods. The simple modal control is not usable in the present case. Indeed, the plant having only 2 inputs, only 2 eigenvalues can be shifted by this method. Now, in the continuous as well as in the discrete case, the integral action adds two unstable poles to the open loop of the augmented system, respectively at $s = 0$ or at $z = 1$. One would thus have during the design process to shift only these two eigenvalues, which has obviously no interest. For similar reasons, the decoupling control is no longer useable either.

b) Study by simulation the influence of load disturbances and of measurement disturbances on the closed-loop response.

Solution:

(a) Run the *MMCE.m* program, with the same plant as previously, by selecting Choice of model: D) discrete, then a synthesis of A) state feedback and Integral action: 1 (= with). Since the program allows partial integral feedback (see Sect. 1.8.3), it prompts the user to indicate, in the form of a row vector, the indices of $\mathbf{y}$ components to feed back by integrators. Answer here: $\begin{pmatrix} 1 & 2 \end{pmatrix}$, i.e. total integral action, on the two measurements.

The program has created the following augmented model (Sect. 1.8.2.3, equation (1.90)):

$$\tilde{\mathbf{\Phi}} = \begin{pmatrix} 0.9679 & 0.0316 & 0.0005 & 0 & 0 \\ 0.0316 & 0.9368 & 0.0313 & 0 & 0 \\ 0.0005 & 0.0313 & 0.9495 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{pmatrix} ; \; \tilde{\mathbf{\Gamma}} = \begin{pmatrix} 0.0752 & 0 \\ 0.0012 & 0.0012 \\ 0 & 0.0744 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} .$$

Design then the state feedback by pole placement, with the following closed-loop poles: –2, –0.5, –1.0359, –10 and –10, the first three poles being chosen so as to allow comparison with Exercise 1.2. The last two poles imposed to the closed-loop system come from the addition of the two integrators, and their choice was guided here by the only concern not to slow down too much the dynamic behavior of the closed-loop system. We obtain:

$$\mathbf{L}_1 = \begin{pmatrix} 9.1213 & 1.7650 & 0.0312 \\ -0.3734 & 2.9596 & 10.1484 \end{pmatrix}, \; \mathbf{L}_2 = \begin{pmatrix} -0.6869 & 0.0163 \\ 0.3012 & -1.4152 \end{pmatrix},$$

$$\mathbf{M} = \begin{pmatrix} 10.221 & 0.696 \\ 0.889 & 12.097 \end{pmatrix}.$$

Since we are dealing here with a multiple input system, the method of Ackermann does not apply anymore, and the program uses here the *place.m* module. Values delivered by an earlier version of the MATLAB® software might differ slightly from these, due to the evolution of the algorithms over the successive versions. The *place.m* module tries indeed to optimize the control system robustness against parameter variations.

The **M** matrix calculated by the program is no longer indispensable here, the cancellation of the steady-state error being guaranteed by the two integrators. If its action is suppressed in the simulation diagram, the step responses are quite comparable with those obtained in question (b) of Exercise 1.2.

If on the contrary this term is active, a strong acceleration of the transients is observed, but at the cost of important overshoots, of the order of 20%, as illustrated in Fig. 1.16. The gain matrix **M** provides then the closed-loop system with an anticipation correction action (see Sect. 1.8.3).

It is left to the reader to verify that with the other design methods roughly identical results are obtained.

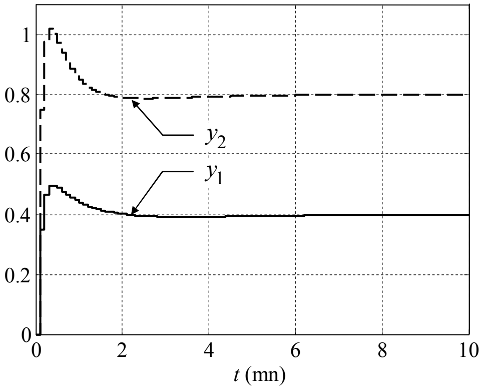


Fig. 1.16 Step responses with feedforward matrix **M**.

(b) The effect of a load disturbance applied at $t = 15$ mn through the vector $\mathbf{e} = (0 \quad 1 \quad 0)^{\mathrm{T}}$ as in question (**b**) of the previous exercise, but with an amplitude $0.5 \times T_s$ m^3/mn in order to produce an effect comparable to the continuous case, is illustrated in the two oscillograms of Fig. 1.17, which have been logged in simulation without anticipation term.

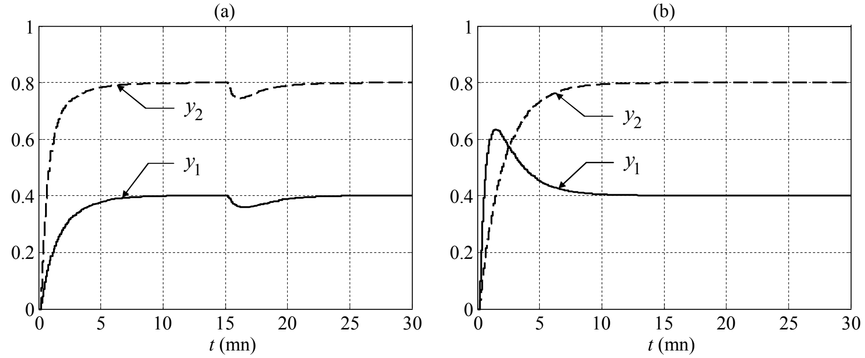


Fig. 1.17 Output step responses, obtained with:
(a) design of question (a); (b) complete modal design, with zero compensation.

The left-hand oscillogram corresponds to the previous synthesis. The right-hand oscillogram has been obtained by performing again a complete modal design and by placing one of the two eigenvalues over the plant zero, thus with the following eigenvalues: –2, –0.5, **–0.664**, –10 and –10, the parameter vectors being chosen as follows: $\begin{pmatrix}1 & 0\end{pmatrix}^{\mathrm{T}}$, $\begin{pmatrix}0 & 1\end{pmatrix}^{\mathrm{T}}$, $\begin{pmatrix}1 & 1\end{pmatrix}^{\mathrm{T}}$, $\begin{pmatrix}1 & -2\end{pmatrix}^{\mathrm{T}}$ and $\begin{pmatrix}-1 & 0\end{pmatrix}^{\mathrm{T}}$. As seen, the disturbance is again completely decoupled from the two outputs.

Exercise 1.5 Lateral Motion Control of a Boeing 747

The equations of the motion of a big airplane, such as a Boeing 747, are separated into two categories, which describe respectively the longitudinal and the lateral motions. It is this last movement which will be the subject of the present exercise.

The lateral motion is composed of rolling (angle φ, angular rate *p*), of yawing (angular rate *r*) and of a side-slip, characterized by the side-slip angle β. Two actuators allow stabilizing this movement and guiding the airplane along the desired trajectory: the angle δa of the ailerons situated at the rear of the wings and the angle δg of the rudder situated in the tail at the plane's rear. It is supposed that there are two sensors, one giving the yaw rate and the other the roll rate.

One of the problems of blown wing aircrafts is to offer a very low damping for one of the lateral motion modes, as will appear below. One of the tasks of the controller to be implemented will be to increase this damping coefficient.

In this exercise the following linearized continuous-time model will be used [5]:

$$\begin{pmatrix}\dot\beta\\ \dot r\\ \dot p\\ \dot\varphi\end{pmatrix}=\begin{pmatrix}-0.0558 & -0.9968 & 0.0802 & 0.0415\\ 0.5980 & -0.1150 & -0.0318 & 0\\ -3.0500 & 0.3880 & -0.4650 & 0\\ 0 & 0.0805 & 1 & 0\end{pmatrix}\begin{pmatrix}\beta\\ r\\ p\\ \varphi\end{pmatrix}+\begin{pmatrix}0.00729 & 0.0583\\ -0.4750 & -2.0100\\ 0.1530 & 0.0241\\ 0 & 0\end{pmatrix}\begin{pmatrix}\delta g\\ \delta a\end{pmatrix},$$

$$\mathbf{y}=\begin{pmatrix}0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\end{pmatrix}\begin{pmatrix}\beta\\ r\\ p\\ \varphi\end{pmatrix}.$$

In open loop, this plant has the following eigenvalues:

- $-0.0329 \pm j\,0.9467$, complex conjugated pair which characterizes the very badly damped yaw motion, also called *Dutch roll*;
- -0.0073, which represents the very slow *spiral mode*;
- -0.5627, which represents the roll mode.

[5] The numerical values of this model are drawn from [FrPE94], p. 686, and have been completed by the author for the second control quantity by reasonable guess.

a) Calculate a state-feedback control law which imposes to the closed loop:

2 eigenvalues $\lambda_{L1}, \lambda_{L2} = -0.5 \pm j1.0$ affected to the side-slip mode, so as to make it faster and better damped than the open-loop Dutch roll mode; choose one of the components of the corresponding eigenvectors $\mathbf{v}_{L1}$ and $\mathbf{v}_{L2}$ such that the modes $e^{\lambda_{L1} t}$ and $e^{\lambda_{L2} t}$ do not appear on the roll angle;

2 eigenvalues $\lambda_{L3} = -2$ and $\lambda_{L4} = -0.5$ in order to accelerate the roll mode and the spiral mode; choose here also the corresponding eigenvectors $\mathbf{v}_{L3}$ and $\mathbf{v}_{L4}$ such that the modes $e^{\lambda_{L3} t}$ and $e^{\lambda_{L4} t}$ do not appear on the side-slip angle.

b) Plot the free response of the side-slip angle and of the roll angle to the following initial conditions: $\beta(0) = 10° = 0.15 \text{ rad}$ and $\varphi(0) = 28° = 0.5 \text{ rad}$.

Solution:

(**a**) Run the *MMCE.m* program, by selecting the fourth plant, lateral motion of the Boeing 747, continuous model, then design a state-feedback control, without integral action, by the complete modal design (Roppenecker's formula), which allows influencing the choice of the closed-loop eigenvectors.

After having opted in this module for the targeted choice of the parameter vectors, allocate the value 0 to the fourth component of the two eigenvectors associated with λ_{L1} and λ_{L2} , and the arbitrary value 1 to one of the two components of the corresponding parameter vectors. Similarly, cancel the first component of the eigenvectors associated with λ_{L3} and λ_{L4} , and give again the arbitrary value 1 to one of the two components of the associated parameter vectors.

This yields the following control law:

$$\mathbf{L} = \begin{pmatrix} -20.1334 & 3.6117 & 13.7122 & 6.7387 \\ 5.0231 & -1.2026 & -3.3337 & -1.6468 \end{pmatrix},$$

and the closed-loop eigenvectors are:

$$\mathbf{V} = \begin{pmatrix} 0.2028 - j0.0414 & 0.2028 + j0.0414 & 0 & 0 \\ 0.0197 - j0.2224 & 0.0197 + j0.2224 & -0.0511 & 0.0595 \\ -0.0016 + j0.0179 & -0.0016 - j0.0179 & -0.3752 & -26.3947 \\ 0 & 0 & 0.1897 & 52.7799 \end{pmatrix}.$$

(b) The free responses are given in Fig. 1.18:

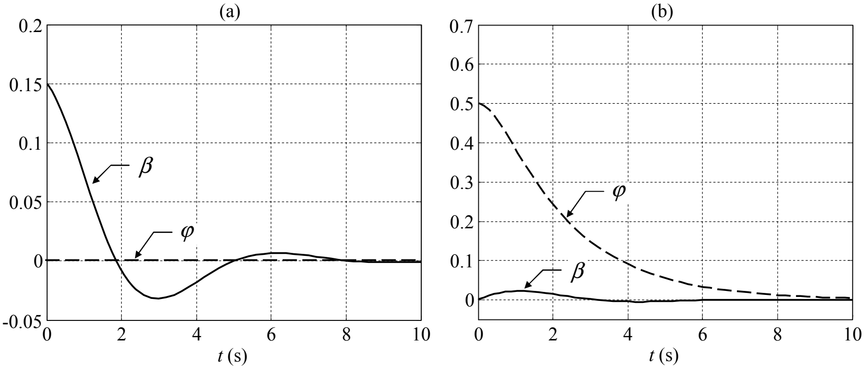


Fig. 1.18 Free responses of: (a) the side-slip angle; (b) the roll angle.

2 Observer Theory

2.1 Introduction to the State Reconstruction Problem

We have seen in the previous chapter that state feedback requires access to the state vector $\mathbf{x}(t)$ or $\mathbf{x}_k$, in its entirety.

However, in many practical cases, even though some state variables will be accessible to measurements, depending on the chosen state representation and the number and layout of the plant sensors, not all states will be measurable.

It is therefore necessary to reconstruct the system state from what is available. Two points of view deserve then consideration.

2.1.1 Purely Deterministic Point of View

The following equations are at our disposal:

Continuous Case: or: *Discrete Case:*

$$\begin{cases}\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \\ \mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u}\end{cases} \qquad \text{or:} \qquad \begin{cases}\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}\mathbf{u}_k \\ \mathbf{y}_k \quad = \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k\end{cases}$$

The question which arises is then the following: why not deduce simply $\mathbf{x}$ from the second equation, $\mathbf{x} = \mathbf{C}^{-1}(\mathbf{y} - \mathbf{D}\mathbf{u})$, since $\mathbf{y}$ and $\mathbf{u}$ are accessible, one as output vector or measurement vector, the other as control vector fed at the input of the plant to be controlled? This is unfortunately impossible, for two reasons:

1. $\mathbf{C}$ is not in general a square matrix $(q \neq n)$, thus is not invertible;
2. if there is additionally measurement noise (see following paragraph), this noise might disturb strongly such a reconstruction.

Since an exact value of $\mathbf{x}(t)$ cannot be obtained, we will content ourselves with imagining a system which, from the known quantities $\mathbf{y}(t)$ and $\mathbf{u}(t)$, will deliver

E. Ostertag, *Mono- and Multivariable Control and Estimation,*
DOI 10.1007/978-3-642-13734-1_2, © Springer-Verlag Berlin Heidelberg 2011

an *approximate value* $\hat{\mathbf{x}}(t)$, or *estimate* of $\mathbf{x}(t)$. Such a system is called *state reconstructor* or *observer*.

The remainder of this chapter deals with the design methods of such observers and their use for state-feedback control or the estimation of disturbances which can be modeled.

2.1.2 Stochastic Context

The previous problem complicates itself if the considered system is submitted to measurement noise and to random disturbances.

The most general form of the state equations of such a system is:

Continuous Case: *Discrete Case:*

$$\begin{cases} \dot{\mathbf{X}} = \mathbf{A}\,\mathbf{X} + \mathbf{B}\mathbf{u} + \mathbf{E}\,\mathbf{V} \\ \mathbf{Y} = \mathbf{C}\mathbf{X} + \mathbf{D}\mathbf{u} + \mathbf{W} \end{cases} \quad \text{or:} \quad \begin{cases} \mathbf{X}_{k+1} = \mathbf{\Phi}\mathbf{X}_k + \mathbf{\Gamma}\mathbf{u}_k + \mathbf{E}\,\mathbf{V}_k \\ \mathbf{Y}_k \quad = \mathbf{C}\,\mathbf{X}_k + \mathbf{D}\mathbf{u}_k + \mathbf{W}_k \end{cases}$$

where: **V** represents the random disturbance vector applied to the state equation;
W represents the random noise vector affecting the measurement.

Note that the system state has become itself a random process $\mathbf{X}(t)$ in the continuous-time case, a random sequence $\mathbf{X}_k$ in the discrete-time case, on account of the presence of a random term in the state equation. Main concepts of random variables or vectors and random processes are summarized in Appendix C.

Remark about notations. We have introduced here capital letters to denote *random* (scalar or vectorial) quantities, to align ourselves on the usual convention of *Probability Theory*, according to which a capital letter such as X (respectively, **X**) denotes a random variable (respectively, vector) and the corresponding lower case letter x (respectively, **x**) a particular deterministic realization of it. Knowing that there will appear no *random matrices* in the rest of this book, there should be no confusion for the reader with the previously adopted notation to represent matrices by bold capital letters.

The state reconstruction problem enters here in the more general frame of what is called *linear filtering*.

Problem of linear filtering. Starting from the previously formulated situation, namely from the plant parameters (**A**, **B**, **C**, **D** or $\mathbf{\Phi}$, $\mathbf{\Gamma}$, **C**, **D**) and from statistical data about the noises **V** and **W** (statistical distribution, mean, variance), the problem consists in finding a causal linear system, at the input of which are applied the signals accessible to measurement, i.e. **u** and **Y**, and which yields at its output a quantity $\widehat{\mathbf{X}}$ as close as possible to the unknown plant state **X**.

Such a system is called a *filter* (Fig. 2.1). The optimal solution of this problem, in the sense of the minimal variance of the estimation error $\mathbf{X}-\widehat{\mathbf{X}}$, is called *Kalman filter*, and will be the subject of Chap. 4.

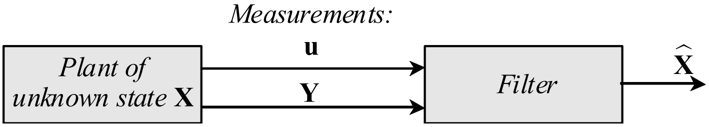


Fig. 2.1 Problem of linear filtering.

2.2 Full-state, or *n*th-order (Luenberger) Observer

2.2.1 Basic Idea of the nth-order Observer

To simplify the mathematical expressions, all the following theory will be developed by assuming $\mathbf{D}=0$, a current situation in practice. One will find, if needed, formulae including the case where $\mathbf{D}\neq 0$ for instance in [LaTh77].

2.2.1.1 Continuous Case

Consider the system

$$\begin{cases}\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{u}\\ \mathbf{y}=\mathbf{C}\mathbf{x}\end{cases}$$

The basic idea of the Luenberger observer [Lue64] [Lue66] [Lue71] consists in adjoining to the above linear time-invariant system another linear time-invariant system, fed permanently by the known signals $\mathbf{u}(t)$ and $\mathbf{y}(t)$, and which must deliver at its output an approximate value $\widehat{\mathbf{x}}(t)$ of $\mathbf{x}(t)$.

The state differential equation of such a system will therefore have the following form, if one assumes, to make this system as simple as possible, that its output matrix **C** is the identity matrix **I**:

$$\dot{\widehat{\mathbf{x}}}=\mathbf{F}\widehat{\mathbf{x}}+\mathbf{J}\,\mathbf{u}+\mathbf{G}\,\mathbf{y}\,, \tag{2.1}$$

where **F**, **J** and **G** are $(n\times n)$, $(n\times p)$ and $(n\times q)$ matrices, to be determined.

While there is no reason to suppose that at $t = 0$ the initial state of the system defined by equation (2.1), $\widehat{\mathbf{x}}_0 = \widehat{\mathbf{x}}(0)$, is the same as the initial state $\mathbf{x}_0$ of the observed plant, it is on the contrary logic to require that $\widehat{\mathbf{x}}(t) \rightarrow \mathbf{x}(t)$ for $t \rightarrow \infty$. Introduce the *estimation error*

$$\tilde{\mathbf{x}}(t) = \mathbf{x}(t) - \widehat{\mathbf{x}}(t) \ . \tag{2.2}$$

$\widehat{\mathbf{x}}(t)$ will be a good *estimate of* $\mathbf{x}(t)$ if $\tilde{\mathbf{x}}(t) \xrightarrow[t\rightarrow\infty]{} 0$.

In order to evaluate the repercussion of that requirement on the choice of the matrices **F**, **J** and **G**, let us establish the differential equation in $\tilde{\mathbf{x}}$:

$$\dot{\tilde{\mathbf{x}}} = \dot{\mathbf{x}} - \dot{\widehat{\mathbf{x}}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} - \mathbf{F}\widehat{\mathbf{x}} - \mathbf{J}\,\mathbf{u} - \mathbf{G}\mathbf{y}\,,$$

or, substituting $\mathbf{y} = \mathbf{C}\,\mathbf{x}$:

$$\dot{\tilde{\mathbf{x}}} = \mathbf{F}\tilde{\mathbf{x}} + (\mathbf{A} - \mathbf{F} - \mathbf{G}\mathbf{C})\mathbf{x} + (\mathbf{B} - \mathbf{J})\mathbf{u}\ .$$

The goal being to obtain that $\tilde{\mathbf{x}}(t) \xrightarrow[t\rightarrow\infty]{} 0$, $\forall\, \mathbf{x}(t)$ and $\mathbf{u}(t)$, one must choose $\mathbf{J} = \mathbf{B}$, and determine **F** and **G** so that $\mathbf{A} - \mathbf{F} - \mathbf{G}\,\mathbf{C} = 0$.

The previous equation becomes then

$$\dot{\tilde{\mathbf{x}}} = \mathbf{F}\,\tilde{\mathbf{x}}\,.$$

This homogeneous differential equation has, according to relation (A.18) of Appendix A, the general solution

$$\tilde{\mathbf{x}}(t) = e^{\mathbf{F}t}\, \tilde{\mathbf{x}}(0)\,.$$

If the eigenvalues of **F** are at the left of the imaginary axis of the complex plane, $\tilde{\mathbf{x}}(t) \rightarrow 0$ for $t \rightarrow \infty$, and the approximate value $\widehat{\mathbf{x}}(t)$ tends towards the true value $\mathbf{x}(t)$.

The system (2.1) defined this way is called a *Luenberger observer*, from its author's name [Lue64].

Definition 2.1. A full-state or Luenberger observer of the system

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\,\mathbf{u} \\ \mathbf{y} = \mathbf{C}\mathbf{x} \end{cases}$$

is a system described by the state differential equation

$$\dot{\hat{\mathbf{x}}} = \mathbf{F}\hat{\mathbf{x}} + \mathbf{B}\mathbf{u} + \mathbf{G}\mathbf{y} \ , \tag{2.3}$$

where the eigenvalues of the **F** matrix are situated in the left-half complex plane and where **G** satisfies the equation:

$$\mathbf{F} = \mathbf{A} - \mathbf{G}\mathbf{C} \ . \tag{2.4}$$

$\hat{\mathbf{x}}(t)$ is called *estimate* of $\mathbf{x}(t)$.

$\tilde{\mathbf{x}}(t) = \mathbf{x}(t) - \hat{\mathbf{x}}(t)$ is the *estimation error*, and $\tilde{\mathbf{x}}(t) \xrightarrow[t\to\infty]{} 0$.

Substituting (2.4) in (2.3), and remembering that $\mathbf{y} = \mathbf{C}\mathbf{x}$, one obtains

$$\dot{\hat{\mathbf{x}}} = \mathbf{A}\hat{\mathbf{x}} + \mathbf{B}\mathbf{u} + \mathbf{G}\mathbf{C}\tilde{\mathbf{x}} \ . \tag{2.5}$$

Block Diagram. Equation (2.5), which is equivalent to (2.3), leads to the state-representation block diagram of Fig. 2.2, where both the system to control and its observer are represented, and where has been introduced the quantity $\tilde{\mathbf{y}} = \mathbf{C}\tilde{\mathbf{x}}$.

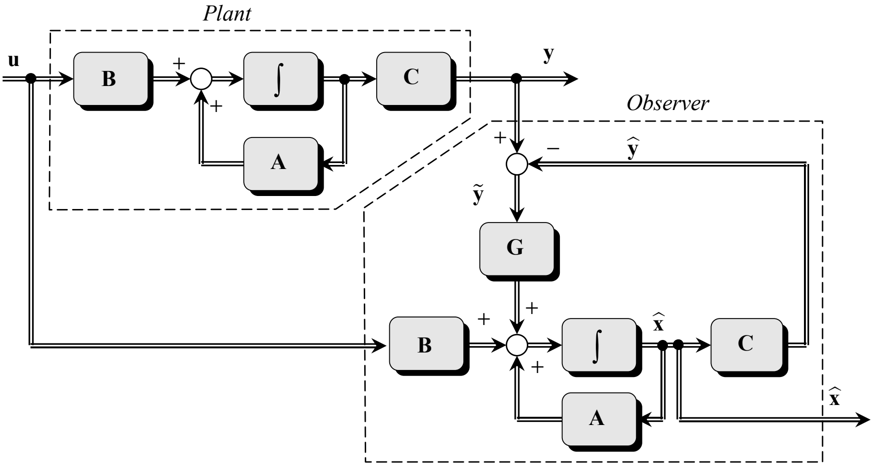
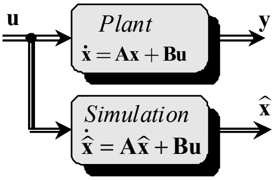


Fig. 2.2 Block diagram of a plant and its observer.

Physical Interpretation. Equation (2.5) suggests the following remarkable interpretation of a Luenberger observer, illustrated in Fig. 2.2: The observer can be seen as a model of the plant, which is driven, not only by the control signal **u** of the latter, but additionally by the estimation error $\tilde{\mathbf{x}}$, in fact by $\mathbf{C}\tilde{\mathbf{x}} = \tilde{\mathbf{y}}$, weighted with a certain factor **G**.

Let us comment this. The Luenberger observer can be considered as being built in two steps:

1°) a simulation of the plant is realized by means of a model, which yields a reconstruction $\widehat{\mathbf{x}}$ of $\mathbf{x}$ as sketched on the right, but which, used like that, has the following drawbacks:


$\mathbf{u}$
Plant
$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$
$\mathbf{y}$
Simulation
$\dot{\widehat{\mathbf{x}}} = \mathbf{A}\widehat{\mathbf{x}} + \mathbf{B}\mathbf{u}$
$\widehat{\mathbf{x}}$

- if the model does not match perfectly the plant, $\widehat{\mathbf{x}}(t)$ will not tend exactly towards $\mathbf{x}(t)$ when $t \rightarrow \infty$;
- if one applies a state feedback from $\widehat{\mathbf{x}}$, namely $\mathbf{u} = -\mathbf{L}\,\widehat{\mathbf{x}}$, the resulting control is finally an *open-loop* control, and not a feedback control system, since the measurement (output vector **y**) is not used;
- if the plant is unstable in open loop, its simulation model will not deliver any usable signal;

2°) a correction term is thus added, containing **y**, or more precisely proportional to the estimation error, $\mathbf{C}\widetilde{\mathbf{x}} = \mathbf{C}\mathbf{x} - \mathbf{C}\widehat{\mathbf{x}} = \mathbf{y} - \widehat{\mathbf{y}} = \widetilde{\mathbf{y}}$, i.e. to the difference between real measured output **y** and estimated output $\widehat{\mathbf{y}}$.

2.2.1.2 Discrete Case

The reasoning is very similar to the continuous case. Let a discrete plant be described by:

$$\begin{cases} \mathbf{x}_{k+1} = \boldsymbol{\Phi}\,\mathbf{x}_k + \boldsymbol{\Gamma}\,\mathbf{u}_k \\ \mathbf{y}_k \quad = \mathbf{C}\,\mathbf{x}_k \end{cases}$$

We will try to build another linear system, with state equation

$$\widehat{\mathbf{x}}_{k+1} = \mathbf{F}\,\widehat{\mathbf{x}}_k + \mathbf{J}\,\mathbf{u}_k + \mathbf{G}\,\mathbf{y}_k\,, \tag{2.6}$$

such that the estimation error

$$\widetilde{\mathbf{x}}_k = \mathbf{x}_k - \widehat{\mathbf{x}}_k$$

tends towards zero when $t \rightarrow \infty$.

We write for that purpose the difference state equation which describes the evolution of the estimation error as a function of time:

$$\widetilde{\mathbf{x}}_{k+1} = \mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1} = \boldsymbol{\Phi}\mathbf{x}_k + \boldsymbol{\Gamma}\mathbf{u}_k - \mathbf{F}\,\widehat{\mathbf{x}}_k - \mathbf{J}\,\mathbf{u}_k - \mathbf{G}\mathbf{y}_k\,,$$

thus, with $\mathbf{y}_k = \mathbf{C}\mathbf{x}_k$,

$$\tilde{\mathbf{x}}_{k+1} = \mathbf{F}\tilde{\mathbf{x}}_k + (\mathbf{\Phi} - \mathbf{F} - \mathbf{GC})\mathbf{x}_k + (\mathbf{\Gamma} - \mathbf{J})\mathbf{u}_k \,. \tag{2.7}$$

We wish here again to cancel as many terms as possible in the second member of this equation. By letting $\mathbf{J} = \mathbf{\Gamma}$, and imposing to **F** and **G** to satisfy the relation

$$\mathbf{\Phi} - \mathbf{F} - \mathbf{G}\mathbf{C} = 0 \,, \tag{2.8}$$

equation (2.7) becomes

$$\tilde{\mathbf{x}}_{k+1} = \mathbf{F}\tilde{\mathbf{x}}_k \,.$$

This homogeneous difference equation represents the free response of a discrete-time system with system matrix **F**, and has, according to (A.19) of Appendix A, the solution

$$\tilde{\mathbf{x}}_k = \mathbf{F}^k\,\tilde{\mathbf{x}}(0) \,.$$

If the eigenvalues of **F** are within the unit circle, $\tilde{\mathbf{x}}_k \xrightarrow[k\to\infty]{} 0$.
Equation (2.6) in turn becomes

$$\hat{\mathbf{x}}_{k+1} = \mathbf{F}\,\hat{\mathbf{x}}_k + \mathbf{\Gamma}\mathbf{u}_k + \mathbf{G}\,\mathbf{y}_k \;\;. \tag{2.9}$$

By substituting, according to (2.8),

$$\mathbf{F} = \mathbf{\Phi} - \mathbf{G}\mathbf{C} \;\;, \tag{2.10}$$

(2.6) can also be put in the form

$$\hat{\mathbf{x}}_{k+1} = \mathbf{\Phi}\,\hat{\mathbf{x}}_k + \mathbf{\Gamma}\mathbf{u}_k + \mathbf{GC}\tilde{\mathbf{x}}_k \;\;. \tag{2.11}$$

The similarity with the continuous case formulae is obvious.

The block diagram is thus the same as the one of the continuous case, where **A** is replaced by $\mathbf{\Phi}$, **B** by $\mathbf{\Gamma}$ and the integrator block by a one sampling-period delay element.

2.2.2 Synthesis of the nth-order Observer

The design occurs in two steps:

1. the eigenvalues of the **F** matrix are chosen, in the left-half complex plane in the continuous case, respectively inside the unit circle in the discrete case, generally to the left of those which have been prescribed to the closed-loop system, i.e. those of $\mathbf{A}-\mathbf{BL}$, respectively closer to the center of the unit circle than those prescribed for $\mathbf{\Phi}-\mathbf{\Gamma L}$, so that the transient response of the observer decays faster than that of the closed-loop system; these eigenvalues should however not be chosen too fast, the observer having then the tendency to amplify high frequency noise, as a result of an excessive bandwidth;
2. the matrix **G** is determined so that $\mathbf{F}=\mathbf{A}-\mathbf{GC}$, respectively $\mathbf{F}=\mathbf{\Phi}-\mathbf{G\,C}$, has indeed these eigenvalues. Let $f(s)$, respectively $f(z)$, denote the corresponding characteristic polynomial, i.e. the polynomial whose roots are these eigenvalues:

$$f(s)=s^n+f_{n-1}\,s^{n-1}+\;\cdots\;+f_1\,s+f_0\,,$$

or

$$f(z)=z^n+f_{n-1}\,z^{n-1}+\;\cdots\;+f_1\,z+f_0\,.$$

To simplify, we will continue the reasoning in the continuous case only. We must thus find a matrix **G** such that

$$\det(s\mathbf{I}-\mathbf{F})=f(s)\,,$$

thus

$$\det\left[s\mathbf{I}-(\mathbf{A}-\mathbf{GC})\right]=f(s)\,. \tag{2.12}$$

Since $\det\mathbf{M}=\det\mathbf{M}^{\mathrm{T}}$ holds for any matrix **M**, (2.12) is equivalent to

$$\det\left[s\mathbf{I}-(\mathbf{A}^{\mathrm{T}}-\mathbf{C}^{\mathrm{T}}\mathbf{G}^{\mathrm{T}})\right]=f(s)\,. \tag{2.13}$$

Formulated this way, the problem is quite similar to an already discussed one, the problem of a state-feedback design by pole placement. It is enough, indeed, to apply the following correspondence:

$$\begin{array}{ccc} \textit{State-feedback Synthesis:} & & \textit{Observer Synthesis:} \end{array}$$

$$\left.\begin{array}{ccc} p(s) & \longrightarrow & f(s) \\ \mathbf{A} & \longrightarrow & \mathbf{A}^{\mathrm{T}} \\ \mathbf{B} & \longrightarrow & \mathbf{C}^{\mathrm{T}} \\ \mathbf{L} & \longrightarrow & \mathbf{G}^{\mathrm{T}} \end{array}\right\}\,, \tag{2.14}$$

to recognize that (2.13) is formally identical to

$$\det\left[s\mathbf{I}-(\mathbf{A}-\mathbf{B}\,\mathbf{L})\right]=p(s).$$

Remark further that (2.14) induces as a corollary the following correspondence:

$$\begin{pmatrix}\mathbf{B} & \mathbf{AB} & \cdots & \mathbf{A}^{n-1}\mathbf{B}\end{pmatrix} \longrightarrow \begin{pmatrix}\mathbf{C}^{\mathrm{T}} & \mathbf{A}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}} & \cdots & (\mathbf{A}^{\mathrm{T}})^{n-1}\mathbf{C}^{\mathrm{T}}\end{pmatrix},$$

which can be written, since $(\mathbf{M}^{\mathrm{T}})^k=(\mathbf{M}^k)^{\mathrm{T}}$ for any matrix **M,**

$$\begin{pmatrix}\mathbf{B} & \mathbf{AB} & \cdots & \mathbf{A}^{n-1}\mathbf{B}\end{pmatrix} \longrightarrow \begin{pmatrix}\mathbf{C}\\ \mathbf{CA}\\ \vdots\\ \mathbf{CA}^{n-1}\end{pmatrix}^{\mathrm{T}}. \tag{2.15}$$

This spells literally as: $\begin{pmatrix}\textit{controllability}\\ \textit{matrix}\ \mathbf{Q}_C\end{pmatrix} \longrightarrow \begin{pmatrix}\textit{observability}\\ \textit{matrix}\ \mathbf{Q}_O\end{pmatrix}^{\mathrm{T}}.$

It had been shown in Chap. 1 that the pole placement problem has always a solution if the plant is controllable. From the last correspondence we deduce therefore that the observer synthesis problem is solvable if the system $\left\{\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{u},\ \mathbf{y}=\mathbf{C}\mathbf{x}\right\}$ is observable. The observer synthesis is a *dual* problem of the controller synthesis in the duality exposed above, due to Kalman [Kal60]. For a mathematical or a practical observability criterion, see Appendix A.

2.2.3 Observer Determination for Single-output Systems

This duality shows that the design problem considered here has a unique solution if $q=1$, thus if the system has only one output:

$$\begin{cases}\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{u}\\ y=\mathbf{c}^{\mathrm{T}}\mathbf{x}\end{cases}$$

The observer representation becomes, instead of (2.3) and (2.4):

$$\dot{\hat{\mathbf{x}}}=\mathbf{F}\hat{\mathbf{x}}+\mathbf{B}\mathbf{u}+\mathbf{g}\,y,\quad \text{where}\quad \mathbf{F}=\mathbf{A}-\mathbf{g}\mathbf{c}^{\mathrm{T}}. \tag{2.16}$$

The present problem is dual to the design of a state-feedback control for a *single-input* system. By adapting the previous duality to this scalar situation and completing the correspondence (2.15) by

$$\begin{aligned} \text{last row of } \mathbf{Q}_C^{-1} &\longrightarrow \text{ last row of } \left(\mathbf{Q}_O^{\mathrm{T}}\right)^{-1} = \left[\text{last column of } \mathbf{Q}_O^{-1}\right]^{\mathrm{T}}, \\ \mathbf{q}_C^{\mathrm{T}} &\longrightarrow \mathbf{q}_O^{\mathrm{T}}, \end{aligned} \tag{2.17}$$

it will be possible to transpose directly by duality the results of Chap. 1.

2.2.3.1 Plants in Observability Canonical Form

Theorem 2.1. *If one assumes that the plant*

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\,\mathbf{u} \\ y = \mathbf{c}^{\mathrm{T}}\mathbf{x} \end{cases}$$

is given in observability canonical form, with the characteristic polynomial

$$s^n + a_{n-1}\,s^{n-1} + \cdots + a_1\,s + a_0\,,$$

and if one prescribes for the Luenberger observer eigenvalues of $\mathbf{F} = \mathbf{A} - \mathbf{g}\,\mathbf{c}^{\mathrm{T}}$ *such that the characteristic polynomial is*

$$s^n + f_{n-1}\,s^{n-1} + \cdots + f_1\,s + f_0\,,$$

the vector $\mathbf{g}$ *must take the following value:*

$$\mathbf{g} = \begin{pmatrix} f_0 - a_0 \\ f_1 - a_1 \\ \vdots \\ f_{n-1} - a_{n-1} \end{pmatrix}. \tag{2.18}$$

Let us write in this case the observer equations explicitly. Since the plant to control is then represented by the following matrices (see relations (A.3) or (A.34) of Appendix A):

$$\mathbf{A} = \begin{pmatrix} 0 & \cdots & \cdots & -a_0 \\ 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & -a_{n-1} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{n-1} \end{pmatrix}, \quad \mathbf{c}^{\mathrm{T}} = \begin{pmatrix} 0 & \cdots & 0 & 1 \end{pmatrix},$$

we have

$$\mathbf{F} = \mathbf{A} - \mathbf{g}\mathbf{c}^{\mathrm{T}} = \mathbf{A} - \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_n \end{pmatrix} \begin{pmatrix} 0 & \cdots & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & \cdots & \cdots & -a_0 - g_1 \\ 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & -a_{n-1} - g_n \end{pmatrix},$$

and, with (2.18),

$$\mathbf{F} = \begin{pmatrix} 0 & \cdots & \cdots & -f_0 \\ 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & -f_{n-1} \end{pmatrix}. \tag{2.19}$$

The observer itself is thus represented in observability canonical form.

2.2.3.2 Plants in Arbitrary State Representation: Use of Ackermann's Formula

By means of the correspondences (2.14) and (2.17) we deduce from the relation (1.24) the following dual formula:

$$\mathbf{g}^{\mathrm{T}} = f_0\, \mathbf{q}_{\mathcal{O}}^{\mathrm{T}} + f_1\, \mathbf{q}_{\mathcal{O}}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} + \cdots + f_{n-1}\, \mathbf{q}_{\mathcal{O}}^{\mathrm{T}} (\mathbf{A}^{\mathrm{T}})^{n-1} + \mathbf{q}_{\mathcal{O}}^{\mathrm{T}} (\mathbf{A}^{\mathrm{T}})^{n},$$

which leads, after transposition, to the statement of a dual theorem of Theorem 1.2 (Ackermann's theorem):

Theorem 2.2. *If the plant* $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$ *is observable, and if one desires that the nth-order (Luenberger) observer has the characteristic polynomial*

$$f(s) = s^n + f_{n-1}\, s^{n-1} + \cdots + f_1\, s + f_0,$$

one must choose

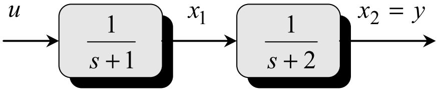
$$\begin{aligned} \mathbf{g} &= f_0\, \mathbf{q}_{\mathcal{O}} + f_1\, \mathbf{A}\mathbf{q}_{\mathcal{O}} + \cdots + f_{n-1}\, \mathbf{A}^{n-1}\, \mathbf{q}_{\mathcal{O}} + \mathbf{A}^n\, \mathbf{q}_{\mathcal{O}} \\ &= f(\mathbf{A})\, \mathbf{q}_{\mathcal{O}} \end{aligned} \tag{2.20}$$

where $\mathbf{q}_{\mathcal{O}}$ *is the last column of the observability matrix inverse,* $\mathbf{Q}_{\mathcal{O}}^{-1}$, *and is determined by*

$$\begin{cases} \mathbf{c}^{\mathrm{T}}\mathbf{q}_{\mathcal{O}} = 0 \\ \vdots \\ \mathbf{c}^{\mathrm{T}}\mathbf{A}^{n-2}\mathbf{q}_{\mathcal{O}} = 0 \\ \mathbf{c}^{\mathrm{T}}\mathbf{A}^{n-1}\mathbf{q}_{\mathcal{O}} = 1 \end{cases} \qquad \textit{or} \qquad \mathbf{q}_{\mathcal{O}} = \mathbf{Q}_{\mathcal{O}}^{-1}\begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

2.2.3.3 Example

Consider the plant represented on the right. Build a full-state observer for this plant, having a double eigenvalue $s = -3$.



The plant state representation is established by mere inspection:

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}u \\ y = \mathbf{c}^{\mathrm{T}}\mathbf{x} \end{cases} \qquad \mathbf{A} = \begin{pmatrix} -1 & 0 \\ 1 & -2 \end{pmatrix}, \ \mathbf{b} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \ \mathbf{c}^{\mathrm{T}} = \begin{pmatrix} 0 & 1 \end{pmatrix},$$

$$\mathbf{Q}_{\mathcal{O}} = \begin{pmatrix} \mathbf{c}^{\mathrm{T}} \\ \mathbf{c}^{\mathrm{T}}\mathbf{A} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -2 \end{pmatrix} \Rightarrow \mathbf{Q}_{\mathcal{O}}^{-1} = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix} \Rightarrow \mathbf{q}_{\mathcal{O}} = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

$$(s+3)^2 = f(s) = s^2 + 6s + 9 \Rightarrow f_1 = 6, \ f_0 = 9,$$

$$\mathbf{A}\mathbf{q}_{\mathcal{O}} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \ \mathbf{A}^2\mathbf{q}_{\mathcal{O}} = \mathbf{A}\cdot\mathbf{A}\mathbf{q}_{\mathcal{O}} = \begin{pmatrix} 1 \\ -3 \end{pmatrix}.$$

Applying (2.20) yields

$$\mathbf{g} = 9\begin{pmatrix} 1 \\ 0 \end{pmatrix} + 6\begin{pmatrix} -1 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ -3 \end{pmatrix} = \begin{pmatrix} 4 \\ 3 \end{pmatrix},$$

then, with (2.16),

$$\mathbf{F} = \mathbf{A} - \mathbf{g}\mathbf{c}^{\mathrm{T}} = \begin{pmatrix} -1 & 0 \\ 1 & -2 \end{pmatrix} - \begin{pmatrix} 4 \\ 3 \end{pmatrix}\begin{pmatrix} 0 & 1 \end{pmatrix} = \begin{pmatrix} -1 & -4 \\ 1 & -5 \end{pmatrix}.$$

The resulting observer has finally the following equation:

$$\dot{\widehat{\mathbf{x}}} = \begin{pmatrix} -1 & -4 \\ 1 & -5 \end{pmatrix}\widehat{\mathbf{x}} + \begin{pmatrix} 1 \\ 0 \end{pmatrix}u + \begin{pmatrix} 4 \\ 3 \end{pmatrix}y.$$

2.3 Reduced-order Observer

2.3.1 Basic Idea

The *n*th-order observer studied previously, and aimed at observing a system which is also of *n*th order, has some degree of redundancy. This results from the fact that this observer builds an estimate of the entire state vector, although part of it is already accessible to the measurement, according to the plant outputs.

Consider again a continuous-time plant described by

$$\begin{cases}\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}\\ \mathbf{y} = \mathbf{C}\mathbf{x}\end{cases}$$

The q outputs being linearly independent, which means to say that the output matrix $\mathbf{C}$, of size $(q \times n)$, has rank q, it is always possible to assume, eventually with a change of basis, that this matrix takes the form

$$\mathbf{C} = \left.\left(\begin{array}{c|c} \underbrace{\begin{matrix}0 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & 0\end{matrix}}_{n-q} & \underbrace{\begin{matrix}1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & 1\end{matrix}}_{q}\end{array}\right)\right\} q \quad = \left(\mathbf{0} \mid \mathbf{I}_q\right). \tag{2.21}$$

Indeed, if initially the system is represented differently and has the output equation $\mathbf{y} = \mathbf{C}'\mathbf{x}'$, the change of coordinates

$$\mathbf{x} = \begin{pmatrix}\mathbf{D}\\ \hline \mathbf{C}'\end{pmatrix}\mathbf{x}' = \mathbf{T}\mathbf{x}',$$

where the matrix $\mathbf{D}$, of size $(n-q)\times n$, is chosen such that $\mathbf{T}$ is regular, will lead to the above representation.

2.3.1.1 Partitioning

Let us then rewrite the system state equations, by partitioning them while taking into account the new form of $\mathbf{C}$, thus by writing

$$\mathbf{x}=\left(\begin{array}{c} x_1 \\ \vdots \\ x_{n-q} \\ \hdashline x_{n-q+1} \\ \vdots \\ x_n \end{array}\right)=\left(\begin{array}{c} \mathbf{v} \\ \hdashline \mathbf{y} \end{array}\right)\begin{array}{l} \left.\begin{matrix}\cdots\\ \vdots\end{matrix}\right\} n-q \\ \left.\begin{matrix}\cdots\\ \cdots\end{matrix}\right\} q \end{array}, \tag{2.22}$$

$$\dot{\mathbf{x}}=\left(\begin{array}{c} \dot{\mathbf{v}} \\ \hdashline \dot{\mathbf{y}} \end{array}\right)=\left(\begin{array}{c:c} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \hdashline \underbrace{\mathbf{A}_{21}}_{n-q} & \underbrace{\mathbf{A}_{22}}_{q} \end{array}\right)\left(\begin{array}{c} \mathbf{v} \\ \hdashline \mathbf{y} \end{array}\right)+\left(\begin{array}{c} \mathbf{B}_1 \\ \hdashline \mathbf{B}_2 \end{array}\right)\mathbf{u}\;\begin{array}{l} \left.\begin{matrix}\cdots\\ \vdots\end{matrix}\right\} n-q \\ \left.\begin{matrix}\cdots\\ \cdots\end{matrix}\right\} q \end{array}, \tag{2.23}$$

or

$$\dot{\mathbf{v}}=\mathbf{A}_{11}\,\mathbf{v}+\mathbf{A}_{12}\,\mathbf{y}+\mathbf{B}_1\,\mathbf{u}\,, \tag{2.24}$$

$$\dot{\mathbf{y}}-\mathbf{A}_{22}\,\mathbf{y}-\mathbf{B}_2\,\mathbf{u}=\mathbf{A}_{21}\,\mathbf{v}\,. \tag{2.25}$$

It will thus suffice to estimate part **v** of the state vector **x**, i.e. $(n-q)$ components.

2.3.1.2 Idea of the Reconstruction

Since **y** is accessible to the measurement, $\dot{\mathbf{y}}$ is it also. Since **u** is also measurable, equations (2.24) and (2.25) can be considered respectively as the state differential equation and the measurement equation of a system which would have,

- as state vector, **v**,
- as input vector, $\mathbf{A}_{12}\,\mathbf{y}+\mathbf{B}_1\,\mathbf{u}$,
- and as output (measurement) vector, $\dot{\mathbf{y}}-\mathbf{A}_{22}\,\mathbf{y}-\mathbf{B}_2\,\mathbf{u}$.

A full-state Luenberger observer, thus of order $(n-q)$, is then built for this system. According to (2.3), if we introduce instead of **G** a new gain matrix $\mathbf{G}_r$ of size $(n-q)\times q$, it will have the following differential state equation:

$$\dot{\widehat{\mathbf{v}}}=(\mathbf{A}_{11}-\mathbf{G}_r\,\mathbf{A}_{21})\widehat{\mathbf{v}}+\mathbf{A}_{12}\,\mathbf{y}+\mathbf{B}_1\,\mathbf{u}+\mathbf{G}_r\,(\dot{\mathbf{y}}-\mathbf{A}_{22}\,\mathbf{y}-\mathbf{B}_2\,\mathbf{u})\,.$$

A priori this observer does not appear satisfactory, since it involves the derivative of **y**. To avoid this inconvenient, let us define

$$\mathbf{z}=\widehat{\mathbf{v}}-\mathbf{G}_r\,\mathbf{y}\,. \tag{2.26}$$

2.3.1.3 State Equations of the Reduced-order Observer

Substituting $\widehat{\mathbf{v}}$ from (2.26) in the previous equation, we obtain:

$$\begin{aligned}\dot{\mathbf{z}} &= (\mathbf{A}_{11}-\mathbf{G}_r\mathbf{A}_{21})\mathbf{z}+\left[(\mathbf{A}_{11}-\mathbf{G}_r\mathbf{A}_{21})\mathbf{G}_r+\mathbf{A}_{12}-\mathbf{G}_r\mathbf{A}_{22}\right]\mathbf{y}+(\mathbf{B}_1-\mathbf{G}_r\mathbf{B}_2)\mathbf{u}\\ \widehat{\mathbf{v}} &= \mathbf{z}+\mathbf{G}_r\mathbf{y}\end{aligned} \qquad (2.27)$$

These equations lead to the block diagram in Fig. 2.3:

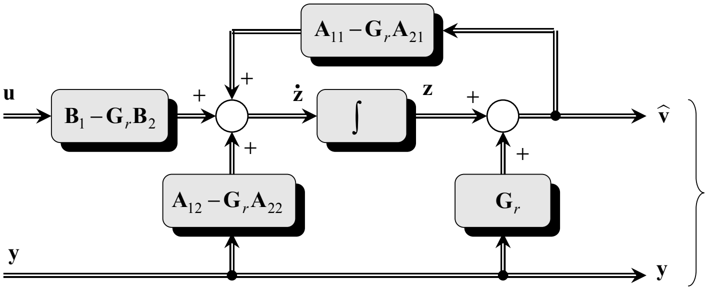

$\mathbf{A}_{11}-\mathbf{G}_r\mathbf{A}_{21}$
u
$\mathbf{B}_1-\mathbf{G}_r\mathbf{B}_2$
+
+
$\dot{\mathbf{z}}$
$\int$
z
+
+
+
$\widehat{\mathbf{v}}$
$\widehat{\mathbf{x}}$
$\mathbf{A}_{12}-\mathbf{G}_r\mathbf{A}_{22}$
$\mathbf{G}_r$
y
y

Fig. 2.3 Block diagram of a reduced-order observer.

Since $\widehat{\mathbf{v}}$, thus also **z**, have the dimension $(n-q)$, this is well a reduced-order observer, of order $(n-q)$.

Lemma. If the pair $(\mathbf{A},\mathbf{C})$ is observable, the pair $(\mathbf{A}_{11},\mathbf{A}_{21})$ is it also. This property stems directly from the definition of observability.

Theorem 2.3. *Given a linear, time-invariant system of order n, which has q linearly independent outputs and is assumed observable, it is possible to construct an observer of order* $(n-q)$ *having arbitrary eigenvalues.*

Remark 2.1. The previous construction yields *one* observer of this type, which has

$$\mathbf{F}=\mathbf{A}_{11}-\mathbf{G}_r\,\mathbf{A}_{21}\,. \qquad (2.28)$$

as system matrix. There are other solutions, as e.g. in the following situation.

Case where $\mathbf{C}=(\mathbf{I}_q \mid \mathbf{0})$ in (2.21). The solution is then obtained by simple permutation of the indices 1 and 2 in (2.27) and (2.28). The corresponding system matrix, for instance, becomes then

$$\mathbf{F}=\mathbf{A}_{22}-\mathbf{G}_r\,\mathbf{A}_{12}\,.$$

Theorem 2.3 claims that it is always possible to reach the fixed objective, provided the hypotheses of the theorem are fulfilled, and this, by imposing arbitrarily the $(n-q)$ eigenvalues of the corresponding $\mathbf{F}$ matrix.

2.3.2 Reduced-order Observer Design for Single-output Systems

Consider again a plant of the form

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \\ y = \mathbf{c}^{\mathrm{T}}\mathbf{x} \end{cases}$$

Now, partition the matrices and vectors as previously, but taking into account the fact that $q=1$. The partition for $\mathbf{x}$ becomes here, according to (2.22):

$$\mathbf{x} = \begin{pmatrix} \mathbf{v} \\ \hline y \end{pmatrix} \begin{matrix} \}\, n-1 \\ \}\, 1 \end{matrix} .$$

It ensues, according to (2.23), that $\mathbf{A}$ and $\mathbf{B}$ are partitioned as follows:

$$\mathbf{A} = \left(\begin{array}{c|c} \mathbf{A}_{11} & \mathbf{a}_{12} \\ \hline \mathbf{a}_{21}^{\mathrm{T}} & a_{22} \end{array}\right) \begin{matrix} \}\, n-1 \\ \}\, 1 \end{matrix} , \quad \mathbf{B} = \left(\begin{array}{c} \mathbf{B}_1 \\ \hline \mathbf{b}_2^{\mathrm{T}} \end{array}\right) \begin{matrix} \}\, n-1 \\ \}\, 1 \end{matrix} , \tag{2.29}$$
$$\qquad \underbrace{\qquad\quad}_{n-1}\ \underbrace{\quad}_{1}$$

and that $\mathbf{c}^{\mathrm{T}}$ must have, according to (2.21), the form $\mathbf{c}^{\mathrm{T}} = \begin{pmatrix} 0 & \cdots & 0 & 1 \end{pmatrix}$, or should be put in it.

The state equations of the reduced-order observer take then, instead of (2.27), the following form, with $\mathbf{G}_r = \mathbf{g}_r$:

$$\begin{aligned} \dot{\mathbf{z}} &= (\mathbf{A}_{11} - \mathbf{g}_r \mathbf{a}_{21}^{\mathrm{T}})\mathbf{z} + \left[(\mathbf{A}_{11} - \mathbf{g}_r \mathbf{a}_{21}^{\mathrm{T}})\mathbf{g}_r + \mathbf{a}_{12} - \mathbf{g}_r a_{22}\right] y + (\mathbf{B}_1 - \mathbf{g}_r \mathbf{b}_2^{\mathrm{T}})\mathbf{u} \\ \hat{\mathbf{v}} &= \mathbf{z} + \mathbf{g}_r y \end{aligned} \tag{2.30}$$

from what results that:

$$\mathbf{F} = \mathbf{A}_{11} - \mathbf{g}_r \mathbf{a}_{21}^{\mathrm{T}} \,. \tag{2.31}$$

The design problem consists thus in prescribing the $(n-1)$ eigenvalues of **F**, which amounts to imposing its characteristic polynomial

$$f(s) = s^{n-1} + f_{n-2}\, s^{n-2} + \cdots + f_1\, s + f_0 \,, \tag{2.32}$$

and determining $\mathbf{g}_r$ such that **F** has effectively this characteristic polynomial.

The above equations can be solved directly by polynomial coefficients identification in the case n is not too large. If this is not the case, one of the following design methods can be used.

2.3.2.1 Case of a SISO Plant in Observability Canonical Form

In this case:

$$\mathbf{A} = \begin{pmatrix} 0 & \cdots & 0 & -a_0 \\ 1 & \ddots & \vdots & \vdots \\ \vdots & \ddots & 0 & \vdots \\ 0 & \cdots & 1 & -a_{n-1} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{n-1} \end{pmatrix}, \quad \mathbf{c}^{\mathrm{T}} = \begin{pmatrix} 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Since $\mathbf{c}^{\mathrm{T}}$ has the form $(\mathbf{0} \mid \mathbf{I}_q)$, no change of coordinates is required. Thus, according to (2.29):

$$\mathbf{A}_{11} = \begin{pmatrix} 0 & \cdots & \cdots & 0 \\ 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{pmatrix}, \quad \mathbf{a}_{12} = \begin{pmatrix} -a_0 \\ \vdots \\ \vdots \\ -a_{n-2} \end{pmatrix}, \qquad \mathbf{B}_1 = \mathbf{b}' = \begin{pmatrix} b_0 \\ \vdots \\ \vdots \\ b_{n-2} \end{pmatrix},$$

$$\mathbf{a}_{21}^{\mathrm{T}} = \begin{pmatrix} 0 & \cdots & 0 & 1 \end{pmatrix}, \quad a_{22} = -a_{n-1}, \qquad \mathbf{b}_2^{\mathrm{T}} = b_{n-1} \,.$$

The posed problem consists then in finding $\mathbf{g}_r = \begin{pmatrix} g_{r,1} & \cdots & g_{r,n-1} \end{pmatrix}^{\mathrm{T}}$ such that **F** given by (2.31) has the characteristic polynomial given by (2.32). If one recognizes the formal similarity of this problem with the one solved by Theorem 2.1, where one would have reduced n by one and made $a_0 = a_1 = \cdots = a_{n-1} = 0$, the solution is obtained readily, according to (2.18):

$$\mathbf{g}_r = \begin{pmatrix} f_0 & \cdots & f_{n-2} \end{pmatrix}^{\mathrm{T}} .$$

Therefore,

$$\mathbf{F} = \begin{pmatrix} 0 & \cdots & 0 & -f_0 \\ 1 & \ddots & \vdots & \vdots \\ \vdots & \ddots & 0 & \vdots \\ 0 & \cdots & 1 & -f_{n-2} \end{pmatrix},$$

by resemblance with (2.19), and finally, by application of (2.30):

$$\begin{aligned} \dot{\mathbf{z}} &= \mathbf{F}\,\mathbf{z} + \left[\mathbf{F}\,\mathbf{g}_r + \mathbf{a}_{12} + a_{n-1}\,\mathbf{g}_r\right] y + (\mathbf{b}' - b_{n-1}\,\mathbf{g}_r)u \\ \widehat{\mathbf{v}} &= \mathbf{z} + \mathbf{g}_r\, y \end{aligned} \tag{2.33}$$

2.3.2.2 Case of an Arbitrary State Representation of the Plant

We apply then the dual Ackermann's formula, Theorem 2.2, formula (2.20), to the reduced system to be observed, of order $(n-q)$, given by the equations (2.24) and (2.25), thus with system matrix $\mathbf{A} = \mathbf{A}_{11}$ and output matrix $\mathbf{c}^{\mathrm{T}} = \mathbf{a}_{21}^{\mathrm{T}}$. The resulting gain vector $\mathbf{g}_r$ is then inserted into (2.30) and (2.31).

2.3.2.3 Example

Consider again the example solved in Sect. 2.2.3.3. The plant being a second order system with one output, it should be possible to build a first order observer having an arbitrary eigenvalue. Let us determine this observer for an eigenvalue $s = -3$.

The plant matrix $\mathbf{c}^{\mathrm{T}} = (0 \quad 1)$ has already the required form. The plant is not here in observability canonical form. We will thus apply successively the direct calculation and the method indicated above (dual Ackerman's formula).

The partitioning according to (2.29) yields here:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad \Rightarrow \quad \begin{cases} a_{11} = -1\,, a_{12} = 0 \\ a_{21} = 1\,, a_{22} = -2 \end{cases}; \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \quad \Rightarrow \quad b_1 = 1\,,\; b_2 = 0\,,$$

Here, $\mathbf{g}_r$ becomes a simple scalar g_r. Applying (2.31) we can thus write:

$$\mathbf{F} = a_{11} - g_r a_{21} = -1 - g_r\,,$$

$$\det(s\mathbf{I} - \mathbf{F}) = s + 1 + g_r = f(s) = s + 3\,.$$

Therefore:

$$g_r = 2\,,$$
$$\mathbf{F} = -1 - 2 = -3\,.$$

Moreover:

$$b_1 - g_r b_2 = 1 - 0 = 1\,,$$
$$a_{12} - g_r a_{22} = 0 + 2g_r = 4\,.$$

Finally, the reduced-order observer has, according to (2.30), the expression:

$$\begin{cases} \dot{z} = -3z - 2y + u \\ \widehat{x}_1 = z + 2y \end{cases}$$

These equations lead to the block diagram of the observer in Fig. 2.4, connected to the plant:

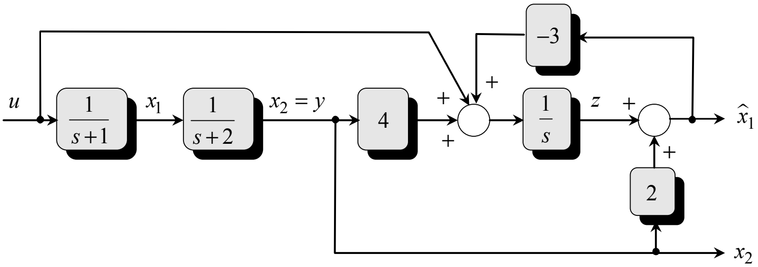


Fig. 2.4 Block diagram of a first-order observer for a second-order plant.

The dual Ackermann's formula applies to the reduced system, thus here to the pair $(\mathbf{c}^{\mathrm{T}}, \mathbf{A})$, with $\mathbf{c}^{\mathrm{T}} = a_{21} = 1$ and $\mathbf{A} = a_{11} = -1$. We obtain thus, successively,

$$\mathbf{Q}_O = \mathbf{c}^{\mathrm{T}} = 1 \quad \Rightarrow \quad \mathbf{q}_O = 1\,,$$

and, with $f(s) = s + 3$,

$$\mathbf{g}_r = f_0 \mathbf{q}_O + f_1 \mathbf{A} \mathbf{q}_O = 3 - 1 = 2\,.$$

Finally, with (2.31) and (2.33), $\mathbf{F} = \mathbf{A} - \mathbf{g}_r \mathbf{c}^{\mathrm{T}} = -1 - 2 = -3$ and

$$\begin{cases} \dot{z} = -3z - 2y + u \\ \widehat{v} = \widehat{x}_1 = z + 2y \end{cases}$$

This result is obviously identical to the previous one.

2.4 Generalized Observer

The previous concepts of *n*th-order observer and reduced-order observer can be grouped under one single theory, thanks to the generalization which will be presented hereafter, and which will additionally allow introducing a new form of reconstructor: the functional observer.

2.4.1 Generalized Observer Definition

To observe a linear time-invariant system, assumed observable, of order n and state vector **x**, another linear time-invariant system can be introduced, of order r and state vector **z**, and having the following state equation [CeBa84]:

$$\dot{\mathbf{z}} = \mathbf{F}\mathbf{z} + \mathbf{J}\,\mathbf{u} + \mathbf{G}\mathbf{y}, \quad \text{where } \operatorname{size}(\mathbf{z}) = (r \times 1)\,. \tag{2.34}$$

Such a system is said to be an observer of the first one if, given an arbitrary transformation matrix **T**, of size $(r \times n)$, its state vector **z** represents an estimation of **Tx**, vector of length r:

$$\mathbf{z} = \mathbf{T}\hat{\mathbf{x}}\ , \tag{2.35}$$

i.e.

$$\mathbf{z} \xrightarrow[t \to \infty]{} \mathbf{T}\mathbf{x}\,. \tag{2.36}$$

Introduce the estimation error

$$\boldsymbol{\varepsilon} = \mathbf{T}\mathbf{x} - \mathbf{z} = \mathbf{T}\mathbf{x} - \mathbf{T}\hat{\mathbf{x}}\ . \tag{2.37}$$

It is thus desired that $\boldsymbol{\varepsilon} \xrightarrow[t \to \infty]{} 0$.

Let us establish the state equations of this generalized observer by following the same reasoning as has been done for the *n*th-order observer. Since the estimation error $\boldsymbol{\varepsilon}$ must tend towards zero for $t \to \infty$, it must be possible to write:

$$\dot{\boldsymbol{\varepsilon}} = \mathbf{F}\boldsymbol{\varepsilon}\,, \tag{2.38}$$

or

$$(\mathbf{z}-\mathbf{T}\mathbf{x})^{\bullet}=\mathbf{F}(\mathbf{z}-\mathbf{T}\mathbf{x}),$$

where **F** is a matrix of size $(r\times r)$, the eigenvalues of which will be situated in the left-half complex plane in all cases.

It comes out that

$$\dot{\mathbf{z}}=\mathbf{F}\mathbf{z}+\mathbf{T}\dot{\mathbf{x}}-\mathbf{F}\mathbf{T}\mathbf{x},$$

and, since $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{u}$,

$$\dot{\mathbf{z}}=\mathbf{F}\mathbf{z}+(\mathbf{T}\mathbf{A}-\mathbf{F}\mathbf{T})\mathbf{x}+\mathbf{T}\mathbf{B}\mathbf{u}.$$

The fact that **x** is accessible only in the form $\mathbf{y}=\mathbf{C}\mathbf{x}$ suggests to look for a matrix **G** such that

$$(\mathbf{T}\mathbf{A}-\mathbf{F}\mathbf{T})\mathbf{x}=\mathbf{G}\mathbf{y},$$

thus such that

$$\mathbf{T}\mathbf{A}-\mathbf{F}\mathbf{T}=\mathbf{G}\mathbf{C}\ , \tag{2.39}$$

where **G** is of size $(r\times q)$ since **C** has size $(q\times n)$ and since **F** and **T** have the above mentioned sizes.

The state equation (2.34) becomes then, with $\mathbf{J}=\mathbf{T}\mathbf{B}$,

$$\dot{\mathbf{z}}=\mathbf{F}\mathbf{z}+\mathbf{T}\mathbf{B}\mathbf{u}+\mathbf{G}\mathbf{y}\ . \tag{2.40}$$

The observers of order n and of reduced order are deduced from this general representation as special cases.

2.4.2 Special Case of the Reduced-order Observer

We dispose here of q measurements and want to construct an observer of order $r=n-q$ which reconstructs the $(n-q)$ non measurable components of **x**.

With the previous hypothesis that $\mathbf{C}=\left(\mathbf{0}\ \vdots\ \mathbf{I}_q\right)$, the vector to estimate is (see (2.22))

$$\hat{\mathbf{x}}=\begin{pmatrix}\hat{\mathbf{v}}\\ --\\ \mathbf{y}\end{pmatrix}\begin{matrix}\}\,n-q\\ \\ \}\,q\end{matrix}\ .$$

By rewriting (2.26) the following way:

$$\mathbf{z}=\hat{\mathbf{v}}-\mathbf{G}_r\mathbf{y}=\left(\mathbf{I}_{n-q} \;\vdots\; -\mathbf{G}_r\right)\begin{pmatrix}\hat{\mathbf{v}}\\ \hline \mathbf{y}\end{pmatrix},$$

$\mathbf{z}$ will have indeed the form (2.35), $\mathbf{z}=\mathbf{T}\hat{\mathbf{x}}$, provided we let

$$\mathbf{T}=\left(\mathbf{I}_{n-q} \;\vdots\; -\mathbf{G}_r\right) \; . \tag{2.41}$$

Let us verify whether, with this transformation **T**, the relations (2.27), which had been obtained with the form of **C** recalled above, are effectively regained.
F will be here a matrix of size $(n-q)\times(n-q)$, and **G** will be of size $(n-q)\times q$.
By applying (2.39), we obtain, with the partition defined in (2.23) for **A** and **B**,

$$\left(\mathbf{I}_{n-q} \;\vdots\; -\mathbf{G}_r\right)\left(\begin{array}{c|c}\mathbf{A}_{11} & \mathbf{A}_{12}\\ \hline \mathbf{A}_{21} & \mathbf{A}_{22}\end{array}\right) - \mathbf{F}\left(\mathbf{I}_{n-q} \;\vdots\; -\mathbf{G}_r\right)=\mathbf{G}\left(\mathbf{0} \;\vdots\; \mathbf{I}_q\right),$$

$$\left(\mathbf{A}_{11}-\mathbf{G}_r\mathbf{A}_{21} \;\vdots\; \mathbf{A}_{12}-\mathbf{G}_r\mathbf{A}_{22}\right) - \left(\mathbf{F} \;\vdots\; -\mathbf{F}\mathbf{G}_r\right)=\left(\mathbf{0} \;\vdots\; \mathbf{G}\right),$$

from where it follows readily that

$$\begin{aligned}\mathbf{F}&=\mathbf{A}_{11}-\mathbf{G}_r\mathbf{A}_{21}\;,\\ \mathbf{G}&=\mathbf{F}\mathbf{G}_r+\mathbf{A}_{12}-\mathbf{G}_r\mathbf{A}_{22}\;.\end{aligned}$$

B being also partitioned as in (2.23), we have also

$$\mathbf{T}\mathbf{B}=\left(\mathbf{I}_{n-q} \;\vdots\; -\mathbf{G}_r\right)\begin{pmatrix}\mathbf{B}_1\\ \hline \mathbf{B}_2\end{pmatrix}=\mathbf{B}_1-\mathbf{G}_r\mathbf{B}_2\;. \tag{2.42}$$

By substituting the values of **F**, **G** and **TB** into (2.40), we regain equation (2.27).

Remark 2.2. The reduced-order observer corresponding to the situation where $\mathbf{C}=\left(\mathbf{I}_q \;\vdots\; \mathbf{0}\right)$ is obtained by letting

$$\mathbf{T}=\left(-\mathbf{G}_r \;\vdots\; \mathbf{I}_{n-q}\right). \tag{2.43}$$

2.4.3 Special Case of the nth-order Observer

This case amounts to considering that no measurement at all is available, thus setting $q = 0$ and constructing an observer of order $r = n$.

It seems then that we could limit the second member of (2.40) to its first two terms. The fact that we want however to conserve the term in $\boldsymbol{y}$ confirms the redundant character of such an observer.

The matrix $\mathbf{T}$ is here of size $(n \times n)$. The estimated vector $\mathbf{z}$ has, in this case, the right dimension *n*. If one wants to have $\mathbf{z} = \widehat{\mathbf{x}}$, it is necessary to let

$$\mathbf{T} = \mathbf{I}_n = \mathbf{I}\,. \tag{2.44}$$

This is the reason why the previously described *n*th-order observer is also called *identity observer*. We regain here again the relations of this observer; in particular (2.39) becomes $\mathbf{A} - \mathbf{F} = \mathbf{G}\mathbf{C}$, which is nothing else than (2.4).

2.4.4 Generalized Observer Output Equation

In the previous sections it was considered that the state vector $\mathbf{z}$ of the generalized observer was giving directly the desired estimated quantity. This is why no output equation was needed.

It may be however interesting to enlarge the representation of the generalized observer given by the state equation (2.40), by adjoining to it an output equation and thus a *generalized* output vector $\boldsymbol{\eta}$, of an arbitrary dimension $m \leq n$, as in [LaTh77], [CeBa84]:

$$\boldsymbol{\eta} = \mathbf{H}\,\mathbf{z} + \mathbf{K}\,\mathbf{y}\ , \tag{2.45}$$

where $\text{size}(\mathbf{H}) = (m \times r)$ and $\text{size}(\mathbf{K}) = (m \times q)$.

By using (2.36), the following will thus hold:

$$\lim_{t \to \infty} \boldsymbol{\eta} = \mathbf{H} \lim_{t \to \infty} \mathbf{z} + \mathbf{K}\,\mathbf{y} = \mathbf{H}\mathbf{T}\,\mathbf{x} + \mathbf{K}\mathbf{C}\,\mathbf{x} = (\mathbf{H}\mathbf{T} + \mathbf{K}\mathbf{C})\,\mathbf{x} = \begin{pmatrix} \mathbf{H} & \vdots & \mathbf{K} \end{pmatrix} \begin{pmatrix} \mathbf{T} \\ \hline \mathbf{C} \end{pmatrix} \mathbf{x} \tag{2.46}$$

In the case where the generalized observer output vector must tend, for $t \to \infty$, towards the *state* of the observed system, which is the case of the two types of observers, or state reconstructors, seen previously, it is of course possible to include them into this more general representation, provided we make there $m = n$ and

$$\left(\mathbf{H} \;\vdots\; \mathbf{K}\right)\left(\begin{array}{c}\mathbf{T}\\ \hline \mathbf{C}\end{array}\right) = \mathbf{I}_n\,, \tag{2.47}$$

which will have the consequence that

$$\lim_{t\to\infty} \boldsymbol{\eta} = \mathbf{x}\,.$$

2.4.4.1 Application to the Reduced-order Observer, Case where $\mathbf{C} = \left(\mathbf{0} \;\vdots\; \mathbf{I}_q\right)$

With the use of the expression (2.41) of the matrix **T** which corresponds to this case and with the help of (B.4) of Appendix B, we can rewrite (2.47) as follows, since the matrices $\mathbf{I}_{n-q}$ and $\mathbf{I}_q$ are obviously invertible:

$$\left(\mathbf{H} \;\vdots\; \mathbf{K}\right) = \left(\begin{array}{c}\mathbf{T}\\ \hline \mathbf{C}\end{array}\right)^{-1} = \left(\begin{array}{c|c}\mathbf{I}_{n-q} & -\mathbf{G}_r\\ \hline \mathbf{0} & \mathbf{I}_q\end{array}\right)^{-1} = \left(\begin{array}{c|c}\mathbf{I}_{n-q} & \mathbf{G}_r\\ \hline \mathbf{0} & \mathbf{I}_q\end{array}\right),$$

or, finally:

$$\mathbf{H} = \left(\begin{array}{c}\mathbf{I}_{n-q}\\ \hline \mathbf{0}\end{array}\right) \quad \text{and} \quad \mathbf{K} = \left(\begin{array}{c}\mathbf{G}_r\\ \hline \mathbf{I}_q\end{array}\right). \tag{2.48}$$

2.4.4.2 Application to the Reduced-order Observer, Case where $\mathbf{C} = \left(\mathbf{I}_q \;\vdots\; \mathbf{0}\right)$

By substituting this time (2.43) into (2.47) and using (B.5) of Appendix B, one obtains

$$\left(\mathbf{H} \;\vdots\; \mathbf{K}\right) = \left(\begin{array}{c}\mathbf{T}\\ \hline \mathbf{C}\end{array}\right)^{-1} = \left(\begin{array}{c|c}-\mathbf{G}_r & \mathbf{I}_{n-q}\\ \hline \mathbf{I}_q & \mathbf{0}\end{array}\right)^{-1} = \left(\begin{array}{c|c}\mathbf{0} & \mathbf{I}_q\\ \hline \mathbf{I}_{n-q} & \mathbf{G}_r\end{array}\right),$$

or, finally:

$$\mathbf{H} = \left(\begin{array}{c}\mathbf{0}\\ \hline \mathbf{I}_{n-q}\end{array}\right) \quad \text{and} \quad \mathbf{K} = \left(\begin{array}{c}\mathbf{I}_q\\ \hline \mathbf{G}_r\end{array}\right). \tag{2.49}$$

2.4.4.3 Application to the Identity Observer

With here $\mathbf{T}=\mathbf{I}_n$ according to (2.44), the only way to satisfy (2.47) $\forall\,\mathbf{C}$ is to let

$$\mathbf{H}=\mathbf{I}_n \quad \text{and} \quad \mathbf{K}=\mathbf{0}\ . \tag{2.50}$$

2.4.5 Application: Functional Observer

2.4.5.1 Principle

For some applications, an estimate of a simple linear combination of the state variables $\varepsilon=\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}$ is all what is needed. It is the case, e.g., for a linear control law of a single-input system, $u=-\boldsymbol{\ell}^{\mathrm{T}}\mathbf{x}$.

The following question arises thus: is it possible in such situations to build a less complex observer than if the entire state were to be reconstructed?

We are faced here with the particular case of the output equation (2.45) corresponding to $m=1$, i.e.:

$$\widehat{\varepsilon}=\mathbf{h}^{\mathrm{T}}\mathbf{z}+\mathbf{k}^{\mathrm{T}}\mathbf{y}\ ,$$

where **z** is the observer state vector and **y** the plant output vector.
For $t\to\infty$, $\widehat{\varepsilon}$ must tend towards $\varepsilon=\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}$. Since then $\mathbf{z}\to\mathbf{T}\mathbf{x}$, (2.46) yields

$$\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}=(\mathbf{h}^{\mathrm{T}}\mathbf{T}+\mathbf{k}^{\mathrm{T}}\mathbf{C})\,\mathbf{x}\,,$$

and

$$\mathbf{h}^{\mathrm{T}}\mathbf{T}+\mathbf{k}^{\mathrm{T}}\mathbf{C}=\boldsymbol{\alpha}^{\mathrm{T}}\ . \tag{2.51}$$

The row vectors $\mathbf{h}^{\mathrm{T}}$, $\mathbf{k}^{\mathrm{T}}$ and $\boldsymbol{\alpha}^{\mathrm{T}}$ are respectively of dimensions *r*, *q* and *n*, if one assumes that the observer order is *r* (see Sect. 2.4.1). We are thus faced with a system of *n* equations to determine the $r+q$ unknowns of $\mathbf{h}^{\mathrm{T}}$ and $\mathbf{k}^{\mathrm{T}}$.

A priori, we could be tempted to choose $r+q=n$, thus to design an observer of reduced order $r=n-q$ to estimate ε. The particular interest of the linear combination estimator resides in the following result, established by [Lue71]:

Proposition 2.1. An arbitrary linear combination of the state, namely $\varepsilon=\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}$, can be estimated by means of an observer having $(\nu-1)$ arbitrary eigenvalues, thus of order $(\nu-1)$, where ν is the *observability index* defined as the smallest positive integer for which the following matrix has the rank *n*:

$$\begin{pmatrix} \mathbf{C} \\ \mathbf{CA} \\ \vdots \\ \mathbf{CA}^{\nu-1} \end{pmatrix}.$$

Since, for any completely observable system, $\nu-1 \le n-q$ and since, for many systems, $\nu-1 \ll n-q$, the observation of a linear combination can be much simpler than that of the state vector itself.

The synthesis takes place as follows:

1. a matrix **F** of size $(\nu-1)\times(\nu-1)$ is chosen, with imposed eigenvalues;
2. $\boldsymbol{\alpha}^{\mathrm{T}}$ being imposed (in the case of the state-feedback control of a single-input system, $\boldsymbol{\alpha}^{\mathrm{T}} = -\boldsymbol{\ell}^{\mathrm{T}}$), as many parameters of $\mathbf{h}^{\mathrm{T}}$, $\mathbf{k}^{\mathrm{T}}$ and **T** as possible are determined, such that (2.51), namely $\mathbf{h}^{\mathrm{T}}\mathbf{T}+\mathbf{k}^{\mathrm{T}}\mathbf{C}=\boldsymbol{\alpha}^{\mathrm{T}}$, is satisfied;
3. **T** and **G** are determined such that the relation $\mathbf{TA}-\mathbf{FT}=\mathbf{GC}$ is satisfied.

The observer obtained this way is called *control observer* or *functional observer*, because it estimates a function of the state variables and not these variables themselves. It is then given by the equations:

$$\begin{aligned} \dot{\mathbf{z}} &= \mathbf{F}\,\mathbf{z}+\mathbf{TB}\,\mathbf{u}+\mathbf{G}\,\mathbf{y} \\ \widehat{\varepsilon} &= \mathbf{h}^{\mathrm{T}}\mathbf{z}+\mathbf{k}^{\mathrm{T}}\mathbf{y} \end{aligned} \tag{2.52}$$

2.4.5.2 Example

Consider the 4th order system represented in Fig. 2.5[1].

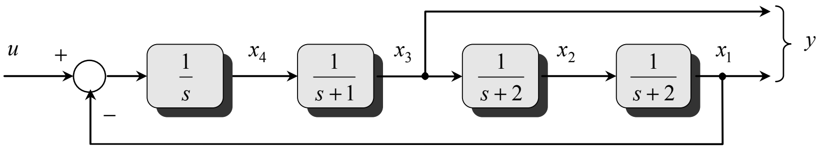


Fig. 2.5 Fourth-order system used in the example.

The available measurements for this system being x_1 and x_3, let us show that it is possible to estimate any linear combination of its state variables by means of a

[1] This example is taken from the paper [Lue71].

functional observer of first order. We will then construct such an observer having a single eigenvalue $s=-3$, which estimates the linear combination x_2+x_4.

The inspection of the functional relations between the four variables, relations all of first order or of pure integrator type, yields directly the following state representation:

$$\begin{cases} \dot{\mathbf{x}} = \begin{pmatrix} -2 & 1 & 0 & 0 \\ 0 & -2 & 1 & 0 \\ 0 & 0 & -1 & 1 \\ -1 & 0 & 0 & 0 \end{pmatrix} \mathbf{x} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} u \\ y = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \mathbf{x} \end{cases}$$

Observability index calculation. Build the matrix composed of the first 2 rows of the observability matrix:

$$\begin{pmatrix} \mathbf{C} \\ \mathbf{CA} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -2 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix} \Rightarrow \det \begin{pmatrix} \mathbf{C} \\ \mathbf{CA} \end{pmatrix} = 1 \neq 0 \Rightarrow \nu = 2 .$$

The observability index of this system being equal to 2, it will thus be possible to calculate a functional observer of first order which estimates a linear combination of the state variables.

Functional observer calculation. From (2.52) we have

$$\begin{cases} \dot{z} = \mathbf{F}z + \mathbf{TB}u + \mathbf{Gy} \\ \widehat{\varepsilon} = \mathbf{h}^{\mathrm{T}} z + \mathbf{k}^{\mathrm{T}} \mathbf{y} = \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{x} = x_2 + x_4 \end{cases} \Rightarrow \left| \begin{array}{l} \text{size}(\mathbf{F}) = (\nu-1)\times(\nu-1) = (1\times 1) \\ \text{size}(\mathbf{T}) = (1\times 4) \\ \text{size}(\mathbf{G}) = (1\times 2) \\ \text{size}(\boldsymbol{\alpha}^{\mathrm{T}}) = (1\times 4) \end{array} \right.$$

Let therefore

$$\mathbf{F} = f\,;\ \mathbf{T} = \begin{pmatrix} t_1 & t_2 & t_3 & t_4 \end{pmatrix};\ \mathbf{G} = \begin{pmatrix} g_1 & g_2 \end{pmatrix};\ \boldsymbol{\alpha}^{\mathrm{T}} = \begin{pmatrix} 0 & 1 & 0 & 1 \end{pmatrix}.$$

The observer characteristic polynomial is

$$\det(s\mathbf{I}-\mathbf{F}) = (s-f) = s+3 \Rightarrow f = -3 .$$

The equation

$$\mathbf{h}^{\mathrm{T}}\mathbf{T}+\mathbf{k}^{\mathrm{T}}\mathbf{C}=\boldsymbol{\alpha}^{\mathrm{T}}$$

becomes, with the present dimensions,

$$h\mathbf{T}+\begin{pmatrix}k_1 & k_2\end{pmatrix}\mathbf{C}=\boldsymbol{\alpha}^{\mathrm{T}},$$

or:

$$h\begin{pmatrix}t_1 & t_2 & t_3 & t_4\end{pmatrix}+\begin{pmatrix}k_1 & 0 & k_2 & 0\end{pmatrix}=\begin{pmatrix}0 & 1 & 0 & 1\end{pmatrix}. \tag{2.53}$$

If we take $h=1$, so as to have for the estimated combination the equivalent of $\widehat{\mathbf{v}}=\mathbf{z}+\mathbf{L}\mathbf{y}$ (see (2.26)), solving (2.53) yields

$$\begin{aligned}
&t_2=t_4=1 \quad\Rightarrow\quad \mathbf{T}=\begin{pmatrix}t_1 & 1 & t_3 & 1\end{pmatrix} \text{ and } \mathbf{TB}=t_4=1,\\
&k_1=-t_1,\\
&k_2=-t_3.
\end{aligned}$$

It remains now to determine t_1, t_3, g_1 and g_2 in such a manner that the equality $\mathbf{TA}-\mathbf{FT}=\mathbf{GC}$ is satisfied. With $\mathbf{GC}=\begin{pmatrix}g_1 & 0 & g_2 & 0\end{pmatrix}$, this condition writes

$$\begin{pmatrix}t_1 & 1 & t_3 & 1\end{pmatrix}\begin{pmatrix}-2 & 1 & 0 & 0\\ 0 & -2 & 1 & 0\\ 0 & 0 & -1 & 1\\ -1 & 0 & 0 & 0\end{pmatrix}+3\begin{pmatrix}t_1 & 1 & t_3 & 1\end{pmatrix}=\begin{pmatrix}g_1 & 0 & g_2 & 0\end{pmatrix},$$

$$\begin{cases}
-2t_1-1+3t_1=g_1,\\
\qquad t_1-2+3=0 \quad\Rightarrow\quad t_1=-1 \quad\Rightarrow\quad g_1=-2 \text{ and } k_1=1;\\
1-t_3+3t_3=g_2,\\
\qquad t_3+3=0 \quad\Rightarrow\quad t_3=-3 \quad\Rightarrow\quad g_2=-5 \text{ and } k_2=3.
\end{cases}$$

The obtained functional observer can be represented by the block diagram of Fig. 2.6. It is as expected a first order dynamic system. A reduced order observer, which would have estimated separately x_2 and x_4, would have been of order 2.

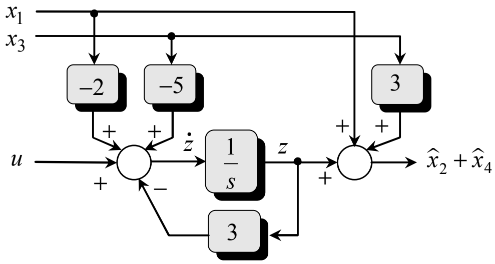


Fig. 2.6 Functional observer estimating a linear combination of 2 states.

2.4.6 Generalized Observer Design Method

A general method to solve (2.39) has been presented in [CeBa84], by using a previous result of [BoYo68]. Due to the complexity of the calculations involved, only the resulting theorem established in this work will be given here without proof.

Theorem 2.4. *If the pair* $(\mathbf{C}, \mathbf{A})$ *is observable and if the spectrum of the matrix* ***F*** *(i.e. the set of its eigenvalues) is distinct from that of* ***A****, the equation* $\mathbf{T\,A}-\mathbf{F\,T}=\mathbf{G\,C}$ *has one unique solution, given by the following formula:*

$$\mathbf{T}=\left(\mathbf{G},\mathbf{FG},\ldots,\mathbf{F}^{r-1}\mathbf{G}\right)\begin{pmatrix}\mathbf{C}\,q_1(\mathbf{A})\\ \mathbf{C}\,q_2(\mathbf{A})\\ \vdots\\ \mathbf{C}\,q_r(\mathbf{A})\end{pmatrix}q_0^{-1}(\mathbf{A})\,, \tag{2.54}$$

where the polynomials $q_i(\lambda)$ *are given by:*

$$\begin{aligned} q_0(\lambda) &= \lambda^r + a_{r-1}\lambda^{r-1} + \ldots + a_1\lambda + a_0 \quad = (\lambda-\lambda_{L1})(\lambda-\lambda_{L2})\ldots(\lambda-\lambda_{Lr}) \\ q_1(\lambda) &= \lambda^{r-1} + a_{r-1}\lambda^{r-2} + \ldots + a_2\lambda + a_1 \\ \vdots\quad & \vdots \qquad\quad \vdots \\ q_i(\lambda) &= \lambda^{r-i} + a_{r-1}\lambda^{r-i-1} + \ldots + a_i \\ \vdots\quad & \vdots \qquad\quad \vdots \\ q_r(\lambda) &= 1 \end{aligned} \tag{2.55}$$

According to [CeBa84] again, the synthesis approach of an observer of order r in its generalized form (2.40), be it for a full-state $(r=n)$ or a reduced order $(r=n-q)$ observer, is then the following:

1. choose an $(r \times r)$ matrix **F** in one of the canonical forms, e.g. the controllability canonical form, which has the r eigenvalues desired for the observer, $\lambda_{O1}, \cdots, \lambda_{Or}$;
2. choose randomly a matrix **G** $(r \times q)$, and deduce from it the matrix **T** $(r \times n)$ which is solution of (2.54);
3. if rank$\begin{pmatrix} \mathbf{T} \\ \mathbf{C} \end{pmatrix} = r + q$, continue, else go back to step 2, and make there a different choice for **G**;
4. the desired observer is then given by (2.40) and (2.45) or one of its variants.

Remark 2.3. In this design, the matrix **G** plays the role of the parameter matrix **P** of the complete modal control (Sect. 1.6).

Remark 2.4. Let us recall here that it is the above theorem which has lead to the establishment by duality of the general design formula of a MIMO state feedback of Theorem 1.7 (Sect. 1.7) [BeOs87].

Remark 2.5. It is clear that the previous approach can also apply to the design of a functional observer. The proof is given in the above cited publication.

2.4.7 Other Design Methods of MIMO Observers

2.4.7.1 Pole Placement Methods and Modal Methods

The Kalman duality highlighted in Sect. 2.2.2 allows obviously using almost all of the state-feedback control design methods seen in Chap. 1 for the design of observers, in the continuous as well as in the discrete case. Let us cite in particular the eigenvalue placement methods for MIMO systems, such as the complete modal synthesis of Sect. 1.6. They allow by duality to select for *all* the eigenvalues of the observer arbitrary values if and only if the plant to be observed is observable.

In the discrete case, it is of course also possible to synthesize a *deadbeat* observer [AsWi97], which has thus all its eigenvalues in the origin of the z-plane, by duality with the design of a control law of this type described in Sect. 1.3. The resulting observer is the fastest possible one, converging to the state of an nth-order discrete system in at most n time steps.

It is enough, indeed, to replace in all those control design algorithms the matrices **A** and **B** respectively by $\mathbf{A}^{\mathrm{T}}$ and $\mathbf{C}^{\mathrm{T}}$, the control matrix **L** obtained in return yielding in this case the transpose of the gain matrix **G** of the full-sate observer or $\mathbf{G}_r$ of the reduced-order observer. This will be illustrated in the solved exercises proposed at the end of this chapter.

2.4.7.2 Application of the Complete Modal Design: Eigenvector Choice

Let us transpose briefly the method described in Sect. 1.6 and consider for simplification purposes the continuous time *n*th-order observer. The discrete case is quite similar.

Once the *n* eigenvalues of the observer system matrix $\mathbf{F}=\mathbf{A}-\mathbf{GC}$, solutions of the equation

$$\det\left[\lambda_{Gi}\,\mathbf{I}-(\mathbf{A}-\mathbf{GC})\right]=0,\quad i=1,\ldots,n\,,$$

have been chosen, the additional degrees of freedom will be used, by introducing this time the *left* eigenvectors $\mathbf{w}_{Gi}$ of **F**, or *right* eigenvectors of $\mathbf{F}^{\mathrm{T}}$, which satisfy the equation

$$\mathbf{w}_{Gi}^{\mathrm{T}}\left[(\mathbf{A}-\mathbf{GC})-\lambda_{Gi}\,\mathbf{I}\right]=0,\quad i=1,\ldots,n\,.$$

The introduction of *n* parameter *row*-vectors $\mathbf{q}_i^{\mathrm{T}}=\mathbf{w}_{Gi}^{\mathrm{T}}\mathbf{G}$, of dimension *q*, changes the previous equation to

$$\mathbf{w}_{Gi}^{\mathrm{T}}(\mathbf{A}-\lambda_{Gi}\mathbf{I})=\mathbf{q}_i^{\mathrm{T}}\,\mathbf{C}\,,$$

which writes also

$$\begin{pmatrix}\mathbf{w}_{Gi}^{\mathrm{T}} & \mathbf{q}_i^{\mathrm{T}}\end{pmatrix}\begin{pmatrix}\mathbf{A}-\lambda_{Gi}\,\mathbf{I}\\ -\mathbf{C}\end{pmatrix}=0\,. \tag{2.56}$$

Equation (2.56), dual of (1.67), represents a system of *n* linear equations in $(n+q)$ unknowns. If the λ_{Gi} are all different from the eigenvalues of **A**, the column matrix in the first member is of full rank equal to *n*, and a unique solution of this equation will be obtained if one chooses arbitrarily the value of *q* unknowns. The arbitrary choice of these *q* unknowns can be spread arbitrarily over the components of $\mathbf{w}_{Gi}$ and those of $\mathbf{q}_i$. It is guided by the following considerations.

The observer transient response, solution of its state equation (2.3) expressed as a function of its eigenmodes, is obtained by replacing, in the formula (A.45) of Appendix A, $\mathbf{x}(t)$ by $\widehat{\mathbf{x}}(t)$ and λ_i by λ_{Gi}, and by taking into account the double input of an observer:

$$\widehat{\mathbf{x}}(t)=\sum_{i=1}^{n}\mathbf{v}_{Gi}\,e^{\lambda_{Gi}\,t}\mathbf{w}_{Gi}^{\mathrm{T}}\widehat{\mathbf{x}}_0+\sum_{i=1}^{n}\mathbf{v}_{Gi}\int_0^t e^{\lambda_{Gi}(t-\tau)}\mathbf{w}_{Gi}^{\mathrm{T}}\left[\mathbf{B}\mathbf{u}(\tau)+\mathbf{G}\mathbf{y}(\tau)\right]d\tau\,.$$

In the light of the discussion at the end of this appendix, it appears that the choice of the left eigenvector $\mathbf{w}_{Gi}$ influences the manner the mode $e^{\lambda_{Gi} t}$ is excited by the various components of $\hat{\mathbf{x}}_0$, also by the various components of the control vector $\mathbf{u}(t)$, through the components of $\mathbf{w}_{Gi}^{\mathrm{T}}\mathbf{B}$, and finally also by the various components of the measurement vector $\mathbf{y}(t)$, through the components of $\mathbf{w}_{Gi}^{\mathrm{T}}\mathbf{G} = \mathbf{q}_i^{\mathrm{T}}$.

An interesting application of these properties may consist in avoiding that a particularly noisy measurement excites the fastest observer dynamics, by imposing to the corresponding component of the parameter vector $\mathbf{q}_i$ to vanish [Duc01].

2.4.7.3 Quadratic Optimization Methods

This transposition by duality can also be applied to the design of a state feedback by quadratic methods, the so-called *optimal control*, which will be discussed in the next chapter, with the aim of designing an *optimal observer*. Since however the real interest of using quadratic optimization methods resides in the possibility to extend their application to the case of noisy systems, the synthesis of such observers, which will become in fact *filters* in this stochastic context (see Sect. 2.1.2), will be presented in Chap. 4 with a completely independent approach.

2.5 State-feedback Control System with Observer

As already mentioned, the matrix $\mathbf{L}$ of a state-feedback $\mathbf{u} = -\mathbf{L}\,\mathbf{x}$ for a control system is calculated first by assuming $\mathbf{x}$ accessible. During this synthesis it is of course paid attention to the stability and a satisfactory dynamic behavior of the closed loop.

The following question arises then: if, as a result of a partially inaccessible $\mathbf{x}$, an observer, i.e. an additional dynamic system, must be introduced in the loop, does the closed-loop system remain stable?

It is important to know the answer to this question, furthermore to know how the introduction of the observer modifies the pole map of the closed-loop system.

2.5.1 State Equations of the Closed-loop Including an Observer

Consider again a linear time-invariant continuous-time plant:

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \\ \mathbf{y} = \mathbf{C}\mathbf{x} \end{cases} \qquad (\mathbf{A}, \mathbf{B}, \mathbf{C}: \text{ constant}).$$

Once equipped with a state feedback, this system will have the following control law, according to (1.2):

$$\mathbf{u} = -\mathbf{L}\mathbf{x} + \mathbf{M}\mathbf{y}_r .$$

Since the input quantity available to the controller is not **x**, which is assumed inaccessible, but only its estimate $\widehat{\mathbf{x}}$ delivered by the observer, the control law becomes

$$\mathbf{u} = -\mathbf{L}\widehat{\mathbf{x}} + \mathbf{M}\mathbf{y}_r , \tag{2.57}$$

as illustrated in Fig. 2.7 in the case of the *n*th-order observer.

As already known, the stability of a control system is an internal property and does not depend on the applied external quantities. To study it, it is therefore enough to consider the regulation behavior alone, thus letting $\mathbf{y}_r = 0$:

$$\mathbf{u} = -\mathbf{L}\widehat{\mathbf{x}} . \tag{2.58}$$

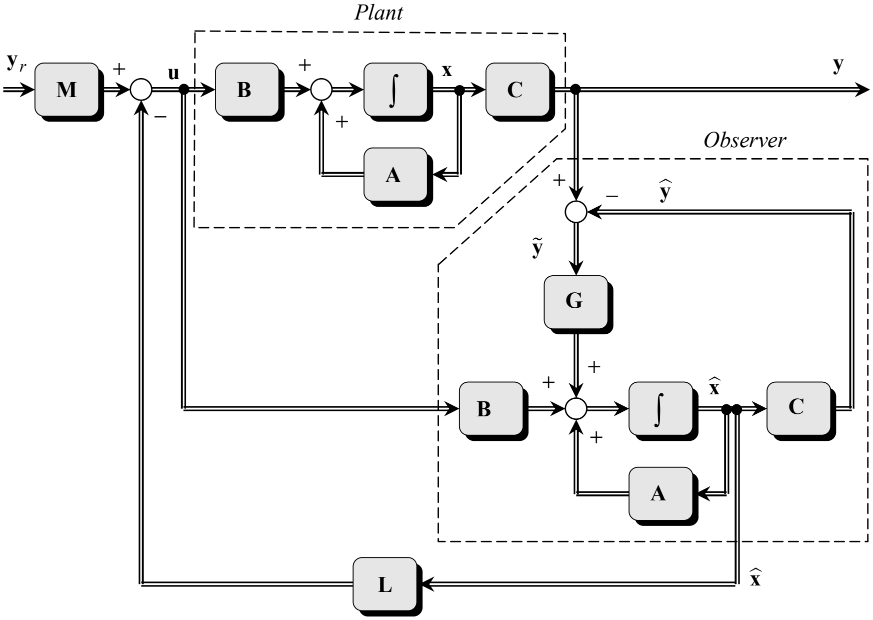


Fig. 2.7 State-feedback control including an *n*-th order observer in the loop.

From the point of view of the observer, considered here in its generalized form described by (2.40) and (2.39), it is clear that **u** must depend on the plant output **y** and on the observer state vector **z** (see (2.45)):

$$\mathbf{u} = \mathbf{H}\,\mathbf{z} + \mathbf{K}\,\mathbf{y}\,,$$

where **H** and **K** are constant matrices.

With $\mathbf{z} = \mathbf{T}\hat{\mathbf{x}}$, according to (2.35), and $\mathbf{y} = \mathbf{C}\,\mathbf{x}$, the two previous equations yield

$$-\mathbf{L}\,\hat{\mathbf{x}} = \mathbf{H}\mathbf{T}\hat{\mathbf{x}} + \mathbf{K}\mathbf{C}\,\mathbf{x}\,.$$

This expression must hold for all values of *t*, thus also for $t \to \infty$, in which case $\hat{\mathbf{x}} = \mathbf{x}$. Therefore

$$-\mathbf{L} = \mathbf{H}\mathbf{T} + \mathbf{K}\mathbf{C}\,, \tag{2.59}$$

equation which generalizes (2.51) encountered in the case of the functional observer.

By taking into account the state differential equation (2.40), the composite system {observer + controller} is described by the equations:

$$\dot{\mathbf{z}} = \mathbf{F}\,\mathbf{z} + \mathbf{T}\mathbf{B}\,\mathbf{u} + \mathbf{G}\,\mathbf{y}\,, \tag{2.60}$$

$$\mathbf{u} = \mathbf{H}\,\mathbf{z} + \mathbf{K}\,\mathbf{y}\,, \tag{2.61}$$

where **F** must be chosen such that its eigenvalues are stable.

The block diagram of the control system is therefore as shown in Fig. 2.8:

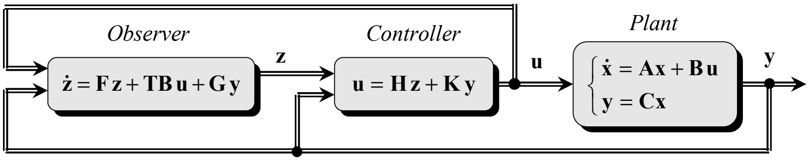

Observer
$\dot{\mathbf{z}} = \mathbf{F}\mathbf{z} + \mathbf{T}\mathbf{B}\mathbf{u} + \mathbf{G}\mathbf{y}$
z
Controller
$\mathbf{u} = \mathbf{H}\mathbf{z} + \mathbf{K}\mathbf{y}$
u
Plant
$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$
$\mathbf{y} = \mathbf{C}\mathbf{x}$
y

Fig. 2.8 Composite system {observer + controller} applied to a plant.

2.5.2 Separation Theorem

The composite dynamic system {plant + observer} is characterized by the composite state vector $\begin{pmatrix}\mathbf{x}^{\mathrm{T}} & \mathbf{z}^{\mathrm{T}}\end{pmatrix}^{\mathrm{T}}$.

The study of the stability of this composite system will be simplified if one introduces, instead of **z**, the estimation error $\boldsymbol{\varepsilon} = \mathbf{T}\mathbf{x} - \mathbf{z}$ already defined previously (see (2.37)). It is solution of the homogeneous differential equation (2.38), recalled here:

$$\dot{\boldsymbol{\varepsilon}} = \mathbf{F}\boldsymbol{\varepsilon} . \tag{2.62}$$

As to the plant state differential equation, it becomes according to (2.61):

$$\begin{aligned}\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} &= \mathbf{A}\mathbf{x} + \mathbf{B}(\mathbf{H}\mathbf{z} + \mathbf{K}\mathbf{y}) \\ &= \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{H}\mathbf{z} + \mathbf{B}\mathbf{K}\mathbf{C}\mathbf{x} ,\end{aligned}$$

and, with the introduction of (2.59):

$$\begin{aligned}\dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} - \mathbf{B}\mathbf{L}\mathbf{x} - \mathbf{B}\mathbf{H}\mathbf{T}\mathbf{x} + \mathbf{B}\mathbf{H}\mathbf{z} \\ &= (\mathbf{A} - \mathbf{B}\mathbf{L})\mathbf{x} - \mathbf{B}\mathbf{H}\boldsymbol{\varepsilon} \qquad .\end{aligned} \tag{2.63}$$

The previous composite system is then described by the two state differential equations (2.62) and (2.63), i.e. by

$$\begin{pmatrix} \dot{\mathbf{x}} \\ \dot{\boldsymbol{\varepsilon}} \end{pmatrix} = \begin{pmatrix} \mathbf{A} - \mathbf{B}\mathbf{L} & -\mathbf{B}\mathbf{H} \\ \mathbf{0} & \mathbf{F} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \boldsymbol{\varepsilon} \end{pmatrix} . \tag{2.64}$$

The characteristic equation of the system closed by means of the observer is thus

$$\begin{vmatrix} \lambda\mathbf{I}_n - (\mathbf{A} - \mathbf{B}\mathbf{L}) & \mathbf{B}\mathbf{H} \\ \mathbf{0} & \lambda\mathbf{I}_r - \mathbf{F} \end{vmatrix} = 0 ,$$

or

$$\left|\lambda\mathbf{I}_n - (\mathbf{A} - \mathbf{B}\mathbf{L})\right| \bullet \left|\lambda\mathbf{I}_r - \mathbf{F}\right| = 0 . \tag{2.65}$$

Since the first of these two determinants yields the eigenvalues of the closed-loop system *without observer* and the second one the observer eigenvalues, the following theorem is obtained:

Theorem 2.5, or Separation Theorem. *The eigenvalues of a control system, closed by means of an observer and containing a plant and a controller both linear and time invariant, are composed of the union of the eigenvalues of the closed-loop system closed without observer and of those of the observer.*

2.5.3 Consequences for the Closed-loop System Stability

A control system, containing an observer in the feedback loop and a linear, time invariant plant and controller, is stable, if the eigenvalues of the closed-loop system closed *without observer* and those of the observer lie in the left-half complex plane.

2.5.4 State-space Design Steps

As already seen in Sect. 1.2.2 for the single input systems, and as has been also proved otherwise for the multiple input systems, the eigenvalues of the closed-loop system without observer can be chosen arbitrarily, provided the plant is controllable [Won67].

Furthermore, as seen in Sect. 2.2 for single output systems and along the subsequent sections for multiple output systems, it is also possible to choose arbitrarily the observer eigenvalues if the plant is observable.

Conclusion. If a linear time-invariant plant is (completely) controllable and observable, it is possible to impose arbitrarily the eigenvalues of the closed-loop system including an observer and a linear time-invariant controller in the loop, as well as the eigenvalues of the observer.

The consequence of Theorem 2.5 is thus, in other words, a complete separation of the two design problems, therefore the name given to it, and the following *Separation Principle*:

1. the controller (matrix **L**) is designed first, without any concern about the eventual need of an observer;
2. if then an observer must be introduced in the loop, this observer will not influence at all the eigenvalues of the closed-loop system just designed in the previous step; it will simply add its own set of eigenvalues to the other ones.

2.5.5 Influence of the Observer Initial State on the System Response

Let us discuss now the total response, i.e. the sum of free and forced responses, of a control system closed by means of an observer, in other words its response to a reference change and to non-zero initial conditions of the plant and of the observer.

The plant is described, in the continuous case, by:

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\,\mathbf{u} \\ \mathbf{y} = \mathbf{C}\mathbf{x} \end{cases} \tag{2.66}$$

and the observer is represented by (2.5), reproduced here:

$$\dot{\widehat{\mathbf{x}}} = \mathbf{A}\widehat{\mathbf{x}} + \mathbf{B}\,\mathbf{u} + \mathbf{G}\mathbf{C}\tilde{\mathbf{x}}\,,$$

$$\dot{\widehat{\mathbf{x}}} = (\mathbf{A} - \mathbf{G}\mathbf{C})\widehat{\mathbf{x}} + \mathbf{B}\,\mathbf{u} + \mathbf{G}\mathbf{C}\,\mathbf{x}\,. \tag{2.67}$$

The control law $\mathbf{u}$ is given by (2.57) (see also Fig. 2.7):

$$\mathbf{u} = -\mathbf{L}\widehat{\mathbf{x}} + \mathbf{M}\,\mathbf{y}_r\,.$$

Let the estimation error appear explicitly in this equation:

$$\mathbf{u} = \mathbf{L}(\mathbf{x} - \widehat{\mathbf{x}}) - \mathbf{L}\,\mathbf{x} + \mathbf{M}\,\mathbf{y}_r\,.$$

The Laplace transform of $\mathbf{u}(t)$, $\mathbf{U}(s) = \mathcal{L}\left[\mathbf{u}(t)\right]$, is then:

$$\mathbf{U}(s) = \mathbf{L}\left[\mathbf{X}(s) - \widehat{\mathbf{X}}(s)\right] - \mathbf{L}\,\mathbf{X}(s) + \mathbf{M}\,\mathbf{Y}_r(s)\,, \tag{2.68}$$

where $\mathbf{X}(s) = \mathcal{L}\left[\mathbf{x}(t)\right]$, $\widehat{\mathbf{X}}(s) = \mathcal{L}\left[\widehat{\mathbf{x}}(t)\right]$ and $\mathbf{Y}_r(s) = \mathcal{L}\left[\mathbf{y}_r(t)\right]$. Apply now the Laplace transform also to the first equation of (2.66), with the use of (2.68):

$$\begin{aligned} s\mathbf{X}(s) &= \mathbf{A}\mathbf{X}(s) + \mathbf{B}\mathbf{U}(s) + \mathbf{x}_0 \\ &= (\mathbf{A} - \mathbf{B}\mathbf{L})\,\mathbf{X}(s) + \mathbf{B}\mathbf{L}\left[\mathbf{X}(s) - \widehat{\mathbf{X}}(s)\right] + \mathbf{B}\mathbf{M}\,\mathbf{Y}_r(s) + \mathbf{x}_0\,. \end{aligned} \tag{2.69}$$

Moreover, the subtraction of (2.67) from the first equation (2.66) yields

$$\dot{\mathbf{x}} - \dot{\widehat{\mathbf{x}}} = (\mathbf{A} - \mathbf{G}\mathbf{C})(\mathbf{x} - \widehat{\mathbf{x}})\,,$$

the Laplace transform of which amounts to

$$s\left[\mathbf{X}(s) - \widehat{\mathbf{X}}(s)\right] = (\mathbf{A} - \mathbf{G}\mathbf{C})\left[\mathbf{X}(s) - \widehat{\mathbf{X}}(s)\right] + (\mathbf{x}_0 - \widehat{\mathbf{x}}_0)\,.$$

Thus

$$(s\mathbf{I} - \mathbf{A} + \mathbf{G}\mathbf{C})\left[\mathbf{X}(s) - \widehat{\mathbf{X}}(s)\right] = \mathbf{x}_0 - \widehat{\mathbf{x}}_0\,.$$

Inserting this result into (2.69) one obtains finally:

$$(s\mathbf{I}-\mathbf{A}+\mathbf{BL})\mathbf{X}(s)=\mathbf{BL}(s\mathbf{I}-\mathbf{A}+\mathbf{GC})^{-1}(\mathbf{x}_0-\widehat{\mathbf{x}}_0)+\mathbf{BM}\,\mathbf{Y}_r(s)+\mathbf{x}_0\,,$$

thus:

$$\begin{aligned}\mathbf{X}(s)=(s\mathbf{I}-\mathbf{A}+\mathbf{BL})^{-1}\mathbf{BM}\,\mathbf{Y}_r(s)+(s\mathbf{I}-\mathbf{A}+\mathbf{BL})^{-1}\mathbf{x}_0+\\+(s\mathbf{I}-\mathbf{A}+\mathbf{BL})^{-1}\mathbf{BL}(s\mathbf{I}-\mathbf{A}+\mathbf{GC})^{-1}(\mathbf{x}_0-\widehat{\mathbf{x}}_0)\ .\end{aligned} \tag{2.70}$$

A close inspection of this formula is instructive in more than one aspect. Consider successively the three second member terms:

- the first term represents the forced response of the closed-loop system to the excitation $\mathbf{y}_r$, with a dynamics dictated by the only eigenvalues of the closed-loop system, which would be closed directly by the state feedback **L** (eigenvalues of the matrix $\mathbf{A}-\mathbf{BL}$); moreover, the relation between $\mathbf{Y}_r(s)$ and $\mathbf{X}(s)$ is exactly the same as if one would apply a direct state feedback, thus the matrix **M** calculated with this hypothesis suits also here;
- the second term represents the free response of the closed-loop system to an initial *plant* state $\mathbf{x}_0$, with a dynamics determined by the same eigenvalues as previously;
- the third term reflects the free response of the closed-loop system to a *difference between the initial values of the plant and the observer*, with a dynamics resulting from both the eigenvalues of the closed-loop which would have been controlled directly by the state feedback **L** and those of the observer (eigenvalues of the matrix $\mathbf{A}-\mathbf{GC}$).

This last point reminds of the separation theorem, but adds the following information: if the observer starts from the *same initial state as the plant*, this third term disappears from the formula; in other words, the observer dynamics is then totally absent from the closed-loop response.

Said differently again, if one could start the algorithm of the observer from the same initial state as the plant $(\widehat{\mathbf{x}}_0=\mathbf{x}_0)$, the closed-loop response would be completely independent from the choice of the poles of this observer, even if they were e.g. *much slower* than those chosen in the design of **L**. The value of $\widehat{\mathbf{x}}(t)$ would in a way be *stuck* to that of $\mathbf{x}(t)$ and evolve with it, in perfect synchronism.

So, when does the observer dynamics make itself feel in the closed-loop dynamic behavior? It does so in two circumstances. First, it is frequent that the plant initial state is not known and that one is obliged to start the observer from an arbitrary initial state, different from that of the plant. The estimated state will then converge towards the real plant state with the observer dynamics only. Second, as soon as a disturbance is applied to the plant, which amounts to apply a new initial state $\mathbf{x}_0 \neq 0$ to it, the observer plays again its role by letting $\widehat{\mathbf{x}}(t)$ converge towards $\mathbf{x}(t)$ with the dynamics corresponding to its eigenvalues.

This mathematical analysis is corroborated by the diagram of the closed-loop system including an observer (Fig. 2.7), where it appears clearly that, if at a given time $\widehat{\mathbf{y}}(t) = \mathbf{y}(t)$, the observer input through its gain matrix **G** is no longer fed by anything. From this time on, the observer behaves thus as a model of the plant, which will consequently evolve with the same dynamics as the plant, for as much as its model matches exactly the plant model.

2.6 Deterministic Disturbances Compensation. Disturbance Observer

In Sect. 1.8 it was shown that constant disturbances applied to the plant could be suppressed by including an integral action in the feedback loop.

In the present section this subject will be extended to disturbances, which are not necessarily constant, but still of deterministic nature. The case of stochastic equation (plant) disturbances or measurement disturbances will be the subject of Chap. 4, devoted to optimal filtering.

Due to the similarity of mathematical developments, only the continuous case will be discussed in details, the final results being then directly transposed to the discrete case. Consider the following continuous time system:

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{E}\mathbf{v} \\ \mathbf{y} = \mathbf{C}\mathbf{x} \end{cases} \tag{2.71}$$

where **v** is a disturbance vector of dimension m influencing the state through a constant $(n \times m)$ matrix **E**, in accordance with the diagram of Sect. 1.8.1.4.

Two situations can then occur, according to whether the disturbance is measurable or not [Föl90].

2.6.1 Case of Measurable Disturbances

The state feedback control law $\mathbf{u} = -\mathbf{L}\,\mathbf{x}$ being determined as seen in Chap. 1, it would be desirable, in order to compensate the effect of the disturbance, to add to it a term $\mathbf{u}_v$ such that the following holds, according to [Föl90]:

$$\mathbf{B}\,\mathbf{u}_v + \mathbf{E}\mathbf{v} = 0\,, \tag{2.72}$$

thus such that $\mathbf{u}_v = -\mathbf{B}^{-1}\mathbf{E}\mathbf{v}$, which would imply of course that **B** be invertible, thus before all that $p = n$. In the infrequent case where this situation is realized, this compensation approach is nothing else than the MIMO extension of the anticipation compensation of the SISO case. The control law $\mathbf{u} = -\mathbf{L}\mathbf{x} + \mathbf{u}_v$ yields then for the closed-loop system the equation

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} - \mathbf{B}\mathbf{L}\mathbf{x} - \mathbf{B}\mathbf{B}^{-1}\mathbf{E}\mathbf{v} + \mathbf{E}\mathbf{v} = (\mathbf{A} - \mathbf{B}\mathbf{L})\mathbf{x} \, ,$$

which confirms the full rejection of the disturbance.

In the more usual case where $p < n$, (2.72) is overdetermined and it is possible to find only an approximate solution in the sense of the least squares, according to the well known method of Gauss. Since the second member cannot be equaled to zero anymore, one attempts merely to make

$$\boldsymbol{\varepsilon} = \mathbf{B}\,\mathbf{u}_v + \mathbf{E}\mathbf{v}$$

as small as possible, in quadratic value:

$$\begin{aligned}
|\boldsymbol{\varepsilon}|^2 = \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon} &= (\mathbf{u}_v^{\mathrm{T}}\mathbf{B}^{\mathrm{T}} + \mathbf{v}^{\mathrm{T}}\mathbf{E}^{\mathrm{T}})(\mathbf{B}\,\mathbf{u}_v + \mathbf{E}\mathbf{v}) \\
&= \mathbf{u}_v^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{B}\,\mathbf{u}_v + \mathbf{u}_v^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{E}\mathbf{v} + \mathbf{v}^{\mathrm{T}}\mathbf{E}^{\mathrm{T}}\mathbf{B}\,\mathbf{u}_v + \mathbf{v}^{\mathrm{T}}\mathbf{E}^{\mathrm{T}}\mathbf{E}\mathbf{v} \\
&= \mathbf{u}_v^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{B}\,\mathbf{u}_v + 2\mathbf{u}_v^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{E}\mathbf{v} + \mathbf{v}_v^{\mathrm{T}}\mathbf{E}^{\mathrm{T}}\mathbf{E}\mathbf{v} \, .
\end{aligned}$$

According to the relations (B.10) and (B.8) of Appendix B,

$$\frac{\partial}{\partial \mathbf{u}_v}|\boldsymbol{\varepsilon}|^2 = 2\mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{u}_v + 2\mathbf{B}^{\mathrm{T}}\mathbf{E}\mathbf{v} \, .$$

B being of full rank, equal to *p*, the matrix $\mathbf{B}^{\mathrm{T}}\mathbf{B}$ of size $(p \times p)$ is regular, and the minimum of $|\boldsymbol{\varepsilon}|^2$ in $\mathbf{u}_v$ will be obtained for

$$\mathbf{u}_v = -(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{E}\;\mathbf{v} = -\mathbf{B}^{\dagger}\mathbf{E}\mathbf{v} \, ,$$

where $\mathbf{B}^{\dagger} = (\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}$ is the pseudoinverse or Moore-Penrose inverse of **B**. The control law is then

$$\mathbf{u} = -\mathbf{L}\mathbf{x} - \mathbf{B}^{\dagger}\mathbf{E}\mathbf{v} \quad \text{or} \quad \mathbf{u} = -\mathbf{L}\hat{\mathbf{x}} - \mathbf{B}^{\dagger}\mathbf{E}\mathbf{v} \, ,$$

the second equation corresponding to the case where an observer is needed to reconstruct the state. It is this last case which is represented in Fig. 2.9, in which appears clearly the anticipation compensation term $\mathbf{B}^{\dagger}\mathbf{E}$:

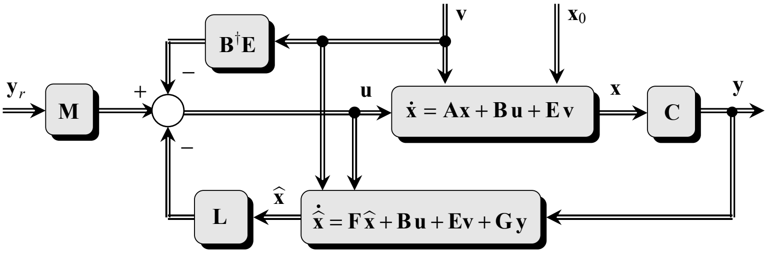

$\mathbf{B}^{\dagger}\mathbf{E}$
$\mathbf{v}$
$\mathbf{x}_0$
$\mathbf{y}_r$
M
+
−
−
u
$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{E}\mathbf{v}$
x
C
y
L
$\widehat{\mathbf{x}}$
$\dot{\widehat{\mathbf{x}}} = \mathbf{F}\widehat{\mathbf{x}} + \mathbf{B}\mathbf{u} + \mathbf{E}\mathbf{v} + \mathbf{G}\mathbf{y}$

Fig. 2.9 Control system with observer and deterministic measurable load disturbance **v**.

Since the disturbance is measurable, it is also injected into the state equation of the observer, which must have the same input signals as the plant. In the figure there is also a static gain compensation matrix **M**, as has been seen in Sect. 1.1.

2.6.2 Case of Non-measurable Disturbances

Though the disturbance is not directly measurable, a knowledge of its form, at least partially, is often available, as is the case with piecewise constant disturbances, the instants of the jumps and their amplitude remaining random, or of sinusoidal signals.

2.6.2.1 Disturbance Model

The disturbance **v** is modeled as the output of a linear system, of state vector $\boldsymbol{\xi}$ and state representation:

$$\begin{cases} \dot{\boldsymbol{\xi}} = \mathbf{A}_{\xi}\boldsymbol{\xi} \\ \mathbf{v} = \mathbf{C}_{\xi}\boldsymbol{\xi} \end{cases} \tag{2.73}$$

with given initial conditions, [AsWi97]. The eigenvalues of $\mathbf{A}_{\xi}$ are commonly located at the origin of the complex plane or on the imaginary axis.

Examples:

- constant disturbance: $\mathbf{A}_{\xi} = 0$;
- undamped sinusoidal disturbance, of angular frequency ω_0 : $\mathbf{A}_{\xi} = \begin{pmatrix} 0 & \omega_0 \\ -\omega_0 & 0 \end{pmatrix}$.

2.6.2.2 Derivation of the Equations in the General Case

This model is then added to the plant state model (2.71) and constitutes with it an *augmented* system, the state representation of which is given by the following equations:

$$\begin{cases} \begin{pmatrix} \dot{\mathbf{x}} \\ \dot{\boldsymbol{\xi}} \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \mathbf{EC}_\xi \\ \mathbf{0} & \mathbf{A}_\xi \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \boldsymbol{\xi} \end{pmatrix} + \begin{pmatrix} \mathbf{B} \\ \mathbf{0} \end{pmatrix} \mathbf{u} \\ \mathbf{y} = \begin{pmatrix} \mathbf{C} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \boldsymbol{\xi} \end{pmatrix} \end{cases} \tag{2.74}$$

This state equation takes the form of a controllability canonical decomposition (see Appendix A, Sect. A.2.3). The state variables $\boldsymbol{\xi}$ are not controllable, which is not a surprise: the disturbance can obviously not be influenced by the control vector **u**. It is thus illusory to try and control this system by an augmented state feedback, which would include the disturbance state vector, $\mathbf{u} = -\mathbf{Lx} - \mathbf{L}_\xi \boldsymbol{\xi}$, even if the plant state **x** is entirely accessible.

One turns then to the estimation of this disturbance by means of a *disturbance observer*, so as to be able to introduce an anticipation compensation based on this estimate, in a way similar to the approach followed in Sect. 2.6.1.

Except in particular cases, the augmented state vector $(\mathbf{x}^{\mathrm{T}} \quad \boldsymbol{\xi}^{\mathrm{T}})^{\mathrm{T}}$ is completely observable, if one assumes that so is the pair (**C A**). It is then possible to synthesize an observer of the augmented system, which will deliver an estimate of both the plant state and the disturbance state, and then to choose a state-feedback control law, which is a linear function of these estimates:

$$\mathbf{u} = -\mathbf{L}\hat{\mathbf{x}} - \mathbf{L}_\xi \hat{\boldsymbol{\xi}} \,. \tag{2.75}$$

The observer state equation is, according to (2.5):

$$\begin{pmatrix} \dot{\hat{\mathbf{x}}} \\ \dot{\hat{\boldsymbol{\xi}}} \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \mathbf{EC}_\xi \\ \mathbf{0} & \mathbf{A}_\xi \end{pmatrix} \begin{pmatrix} \hat{\mathbf{x}} \\ \hat{\boldsymbol{\xi}} \end{pmatrix} + \begin{pmatrix} \mathbf{B} \\ \mathbf{0} \end{pmatrix} \mathbf{u} + \begin{pmatrix} \mathbf{G} \\ \mathbf{G}_\xi \end{pmatrix} \mathbf{C}(\mathbf{x} - \hat{\mathbf{x}}) \,. \tag{2.76}$$

Substituting (2.75) into the first state equation of (2.74), to which the term $-\mathbf{BLx}$ is subtracted, then added, yields for the closed-loop state equation:

$$\dot{\mathbf{x}} = (\mathbf{A} - \mathbf{BL})\mathbf{x} + \mathbf{BL}(\mathbf{x} - \hat{\mathbf{x}}) + \mathbf{EC}_\xi \boldsymbol{\xi} - \mathbf{BL}_\xi \hat{\boldsymbol{\xi}} \,.$$

By introducing the estimation errors of the plant state, $\tilde{\mathbf{x}} = \mathbf{x} - \hat{\mathbf{x}}$, and of the disturbance model state, $\tilde{\boldsymbol{\xi}} = \boldsymbol{\xi} - \hat{\boldsymbol{\xi}}$, the closed-loop equations become:

$$
\begin{aligned}
&\dot{\mathbf{x}} = (\mathbf{A} - \mathbf{BL})\mathbf{x} + (\mathbf{EC}_\xi - \mathbf{BL}_\xi)\boldsymbol{\xi} + \mathbf{BL}\tilde{\mathbf{x}} + \mathbf{BL}_\xi\tilde{\boldsymbol{\xi}} \\
&\dot{\boldsymbol{\xi}} = \mathbf{A}_\xi \boldsymbol{\xi} \\
&\dot{\tilde{\mathbf{x}}} = (\mathbf{A} - \mathbf{LC})\tilde{\mathbf{x}} + \mathbf{EC}_\xi\tilde{\boldsymbol{\xi}} \\
&\dot{\tilde{\boldsymbol{\xi}}} = \mathbf{A}_\xi\tilde{\boldsymbol{\xi}} - \mathbf{G}_\xi\mathbf{C}\tilde{\mathbf{x}}
\end{aligned}
\qquad (2.77)
$$

The state feedback matrix **L** is determined as previously, by one of the methods described in Chap. 1. A judicious choice of the gain matrix $\mathbf{L}_\xi$ allows reducing the effect of the disturbance, in particular if it is possible to satisfy $\mathbf{EC}_\xi - \mathbf{BL}_\xi = 0$.

The block diagram of such a control system is shown in Fig. 2.10.

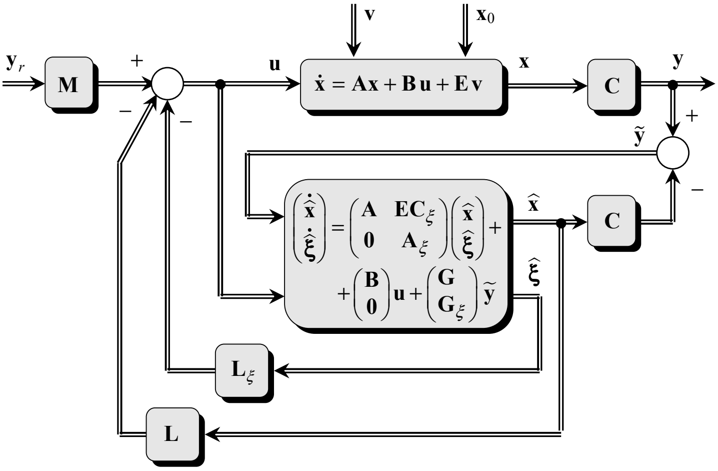

v
x₀
y_r
M
+
−
−
u
ẋ = Ax + Bu + Ev
x
C
y
+
ỹ
−
x̂
C
ξ̂
$\begin{pmatrix}\dot{\hat{\mathbf{x}}}\\ \dot{\hat{\boldsymbol{\xi}}}\end{pmatrix} = \begin{pmatrix}\mathbf{A} & \mathbf{EC}_\xi\\ \mathbf{0} & \mathbf{A}_\xi\end{pmatrix}\begin{pmatrix}\hat{\mathbf{x}}\\ \hat{\boldsymbol{\xi}}\end{pmatrix} + \begin{pmatrix}\mathbf{B}\\ \mathbf{0}\end{pmatrix}\mathbf{u} + \begin{pmatrix}\mathbf{G}\\ \mathbf{G}_\xi\end{pmatrix}\tilde{\mathbf{y}}$
L_ξ
L

Fig. 2.10 Control system with disturbance observer and a non-measurable load disturbance **v**.

2.6.2.3 Particular Case of a Constant Disturbance, Acting at the Plant Input

This case is interesting, since it corresponds to the dry friction of the actuator of a mechanical plant. Such a disturbance is modeled by (2.73) with $\mathbf{A}_\xi = 0$ and $\mathbf{C}_\xi = \mathbf{I}$, thus $\mathbf{v} = \boldsymbol{\xi}$. This disturbance being applied to the plant control input, one has also $\mathbf{E} = \mathbf{B}$. The equations (2.77) are written here:

$$\begin{aligned}
&\dot{\mathbf{x}} = (\mathbf{A} - \mathbf{B}\mathbf{L})\mathbf{x} + \mathbf{B}(\mathbf{I} - \mathbf{L}_\xi)\boldsymbol{\xi} + \mathbf{B}\mathbf{L}\tilde{\mathbf{x}} + \mathbf{B}\mathbf{L}_\xi\tilde{\boldsymbol{\xi}} \\
&\dot{\boldsymbol{\xi}} = \dot{\mathbf{v}} = 0 \\
&\dot{\tilde{\mathbf{x}}} = (\mathbf{A} - \mathbf{L}\mathbf{C})\tilde{\mathbf{x}} + \mathbf{B}\tilde{\boldsymbol{\xi}} \\
&\dot{\tilde{\boldsymbol{\xi}}} = -\mathbf{G}_\xi\mathbf{C}\tilde{\mathbf{x}}
\end{aligned}$$

The disturbance can be eliminated completely from the state equation by choosing $\mathbf{L}_\xi = \mathbf{I}$. The control law (2.75) and the observer state equation (2.76) are then given respectively by

$$\mathbf{u} = -\mathbf{L}\hat{\mathbf{x}} - \mathbf{L}_\xi\hat{\boldsymbol{\xi}} = -\mathbf{L}\hat{\mathbf{x}} - \hat{\mathbf{v}}\ ,$$

and

$$\begin{cases} \dot{\hat{\mathbf{x}}} = \mathbf{A}\hat{\mathbf{x}} + \mathbf{B}(\hat{\mathbf{v}} + \mathbf{u}) + \mathbf{G}\mathbf{C}\tilde{\mathbf{x}} \\ \dot{\hat{\mathbf{v}}} = \dot{\hat{\boldsymbol{\xi}}} = \mathbf{G}_\xi\mathbf{C}\tilde{\mathbf{x}} \end{cases} .$$

Remarking that the second equation can also be written

$$\hat{\mathbf{v}} = \mathbf{G}_\xi\mathbf{C}\int\tilde{\mathbf{x}}(t)dt = \mathbf{G}_\xi\int\tilde{\mathbf{y}}(t)dt\,,$$

the disturbance estimate is proportional to the integral of the state estimation error.

2.6.2.4 Example: Use of a Disturbance Functional Observer for the Control and Simultaneous Dry Friction Cancellation

The present application, which may concern any mechanical system with an actuator submitted to dry friction, has been studied on an inverted pendulum by [BaBO88] and will be handled numerically in the solved exercises at the end of this chapter. It constitutes also a transposition to the discrete-time systems of the material which has been presented previously in the continuous case.

Its principle consists in representing the dry friction force f_c as a constant disturbance, applied to the unique control input u of the plant. Its amplitude is unknown and its sign changes at each change of direction of the cart carrying the inverted pendulum. The state equation of the continuous time plant is thus:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}(u + f_c)\,.$$

The friction is modeled by the equations (2.73), with $\mathbf{A}_\xi = 0$ and $\mathbf{C}_\xi = \mathbf{I}$, and thus also $f_c = \xi$ and $\mathbf{E} = \mathbf{b}$.

An augmented model of the plant and the disturbance is built as in Sect. 2.6.2.2, according to (2.74), and represented by the matrices $\tilde{\mathbf{A}}$, $\tilde{\mathbf{b}}$ and $\tilde{\mathbf{C}}$ defined below:

$$\begin{cases} \begin{pmatrix} \dot{\mathbf{x}} \\ \dot{\xi} \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \mathbf{b} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \xi \end{pmatrix} + \begin{pmatrix} \mathbf{b} \\ 0 \end{pmatrix} u = \tilde{\mathbf{A}} \begin{pmatrix} \mathbf{x} \\ \xi \end{pmatrix} + \tilde{\mathbf{b}} u \\ \mathbf{y} = \begin{pmatrix} \mathbf{C} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \xi \end{pmatrix} = \tilde{\mathbf{C}} \begin{pmatrix} \mathbf{x} \\ \xi \end{pmatrix} \end{cases} \tag{2.78}$$

An observer is designed then for this augmented system. The disturbance step jumps will be interpreted by the observer as sudden changes of the initial conditions of ξ in the integration of the state equation $\dot{\xi} = 0$, and the observer will undergo each time simply a new transient response.

Let us switch now to the discrete model of the plant, sampled at a period T_s. Initial plant:

$$\begin{cases} \mathbf{x}_{k+1} = \mathbf{\Phi} \mathbf{x}_k + \boldsymbol{\gamma}\, u_k \\ \mathbf{y}_k = \mathbf{C} \mathbf{x}_k \end{cases} \quad \text{with } \mathbf{\Phi} = e^{\mathbf{A} T_s};\ \boldsymbol{\gamma} = \int_0^{T_s} e^{\mathbf{A} T_s}\, \mathbf{b}\, dt\,.$$

Augmented model:

$$\begin{cases} \begin{pmatrix} \mathbf{x}_{k+1} \\ \xi_{k+1} \end{pmatrix} = \tilde{\mathbf{\Phi}} \begin{pmatrix} \mathbf{x}_k \\ \xi_k \end{pmatrix} + \tilde{\boldsymbol{\gamma}} u_k \\ y_k = \tilde{\mathbf{C}} \begin{pmatrix} \mathbf{x}_k \\ \xi_k \end{pmatrix} \end{cases} \quad \text{with } \tilde{\mathbf{\Phi}} = e^{\tilde{\mathbf{A}} T_s};\ \tilde{\boldsymbol{\gamma}} = \int_0^{T_s} e^{\tilde{\mathbf{A}} T_s}\, \tilde{\mathbf{b}}\, dt\,.$$

The control law will have the form:

$$u_k = -\boldsymbol{\ell}^{\mathrm{T}} \mathbf{x}_k + M\, y_{r,k} - f_{c,k} = -\begin{pmatrix} \boldsymbol{\ell}^{\mathrm{T}} & 1 \end{pmatrix} \begin{pmatrix} \mathbf{x}_k \\ \xi_k \end{pmatrix} + M\, y_{r,k}\,, \tag{2.79}$$

with $f_{c,k} = \xi_k$. Suppose the plant is of 4th order and only the state variables x_1 and x_2 are measured. Denoting ε_k the part of u_k which is not accessible to the measurement, namely the following linear combination of the state variables x_3 and x_4 and of the disturbance ξ_k:

$$\varepsilon_k = -\begin{pmatrix} 0 & 0 & \ell_3 & \ell_4 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{x}_k \\ \xi_k \end{pmatrix} = -\boldsymbol{\alpha}^{\mathrm{T}} \begin{pmatrix} \mathbf{x}_k \\ \xi_k \end{pmatrix}, \tag{2.80}$$

of which the observer will yield an estimate $\widehat{\varepsilon}_k$, the control law (2.79) becomes:

$$u_k = -\ell_1 x_{1,k} - \ell_2 x_{2,k} + \widehat{\varepsilon}_k + M\, y_{r,k}\,.$$

The functional observer will have the state representation

$$\begin{cases} \mathbf{z}_{k+1} = \mathbf{F}\,\mathbf{z}_k + \mathbf{T}\tilde{\boldsymbol{\gamma}} u_k + \mathbf{g}\, y_k \\ \quad \widehat{\varepsilon}_k = \mathbf{h}^{\mathrm{T}} \mathbf{z}_k + k\, y_k \end{cases}$$

where **F**, **T** and **g** are constant, real matrices and vector, satisfying:

$$\mathbf{T}\tilde{\boldsymbol{\Phi}} - \mathbf{F}\mathbf{T} = \mathbf{g}\tilde{\mathbf{c}}^{\mathrm{T}}\,,$$

and where $\mathbf{h}^{\mathrm{T}}$ and k satisfy: $\mathbf{h}^{\mathrm{T}}\mathbf{T} + k\,\tilde{\mathbf{c}}^{\mathrm{T}} = \boldsymbol{\alpha}^{\mathrm{T}}$.

2.7 Solved Exercises

Some exercises of this chapter are the sequel of exercises which have been solved at the end of Chap. 1. More details can be found there.

Exercise 2.1 Control of an Inverted Pendulum (contd)

Consider once more the inverted pendulum LIP 100 (Amira). We wish this time to compensate the absence of some state variables among the measurements by implementing an observer. This exercise will be solved in continuous time, without integral action. The reader is invited to repeat this solving by using the discrete time model.

a) Repeat the design of a control law, with the same imposed poles as in Exercise 1.1, but by choosing this time another design method than pole placement (“acker.m” or “place.m”), and compare the obtained control law with the previous one.

b) Determine, in the case where only the cart position is measured, a full-state (Luenberger) observer having a unique eigenvalue of order $n = 4$ at $s = -6$.

c) Determine a reduced-order observer in the case where the two measurements mentioned in Exercise 1.1 are available, by imposing again to this observer a unique eigenvalue $s = -6$.

d) Test the control system determined above, with the two observer types successively, by simulation with the Simulink® software, and by selecting different initial states for the plant and the observer.

e) What happens once the observer initial transient response has decayed? Examine this by applying a square signal as a reference in the simulation diagram.

Solution:

(a) Run the *MMCE.m* program, with the plant inverted pendulum LIP100 (Amira), with two measurements, then Choice of model: continuous, Synthesis of state-feedback (default), Integral action: 0 = without (default), Component of y to be set to its reference value: 1. The design methods simple modal control and decoupling method do not apply here, the first because it allows shifting only one single eigenvalue, the second because it applies only to plants having identical number of inputs and outputs. The complete modal synthesis (Roppenecker) could be used, but at the condition to separate slightly the 4 prescribed eigenvalues since this method does not accept a number of identical eigenvalues larger than the number of plant control inputs.

Therefore, select here 5) general formula (Becker-Ostertag), with the controllability canonical form for the closed-loop system and $\lambda_{Li} = -5\,\mathrm{s}^{-1}$, $i = 1, \cdots, 4$. The diagonal canonical form would not be adapted here, since it makes the general formula identical to the Roppenecker's formula and since we want to introduce a multiple eigenvalue of multiplicity order larger than the number of plant inputs (see Remarks of Sect. 1.7.1). Exactly the same control law as the one of Exercise 1.1 is obtained here.

(b) Returning to the Main menu, select the full-state (Luenberger) observer design. The pole placement is not applicable here, since we want to place 4 observer poles, whereas the plant has only 2 outputs. We can on the contrary select again the general formula, with controllability canonical form, which does not have this restriction, and the following parameter vectors: $[1\ \ 1]^{\mathrm{T}}$, $[0\ \ 1]^{\mathrm{T}}$, $[1\ \ 0]^{\mathrm{T}}$ and $[1\ \ -2]^{\mathrm{T}}$. The observer is calculated in the generalized observer form, with

$$\mathbf{G} = \begin{pmatrix} 9.61 & 8.95 \\ 12.65 & 12.34 \\ 17.58 & 18.06 \\ 218.30 & 221.14 \end{pmatrix}, \quad \mathbf{F} = \mathbf{A} - \mathbf{G}\mathbf{C}_m = \begin{pmatrix} -9.61 & -8.95 & -1.950 & 0 \\ -12.65 & -12.34 & 0 & 1 \\ -17.58 & -18.19 & -1.915 & 0.0008 \\ -218.30 & -199.66 & 26.339 & -0.136 \end{pmatrix}.$$

(c) From the Main menu, select this time the reduced-order observer design and the Shape of C: (I_q | 0). It is possible here to synthesize the observer gain matrix $\mathbf{G}_r$ by pole placement since there are only 2 poles to prescribe, the reduced-order observer being of order $n - q = 2$. We get:

$$\mathbf{G}_r = \begin{pmatrix} -2.095 & 0.001 \\ -13.505 & 5.864 \end{pmatrix}, \quad \mathbf{F} = \mathbf{A}_{22} - \mathbf{G}_r \mathbf{A}_{12} = \begin{pmatrix} -6 & 0 \\ 0 & -6 \end{pmatrix},$$

with $\mathbf{J} = \mathbf{TB} = \mathbf{B}_2 - \mathbf{G}_r \mathbf{B}_1$ according to (2.42) and **H** and **K** given by (2.49).

(d) The simulation diagrams proposed in Appendix D allow connecting the observer in the loop or leaving it unused, so that we can study its dynamical behavior independently from that of the control loop, by closing the loop on the state. The free responses from the initial states $\mathbf{x}_0 = \begin{pmatrix} 0 & -2 & 0.5 & 0 \end{pmatrix}^{\mathrm{T}}$ and $\mathbf{z}_0 = 0$, which are given in Fig. 2.11, have been recorded with the two types of observer acting *in free wheel*, thus out of the loop.

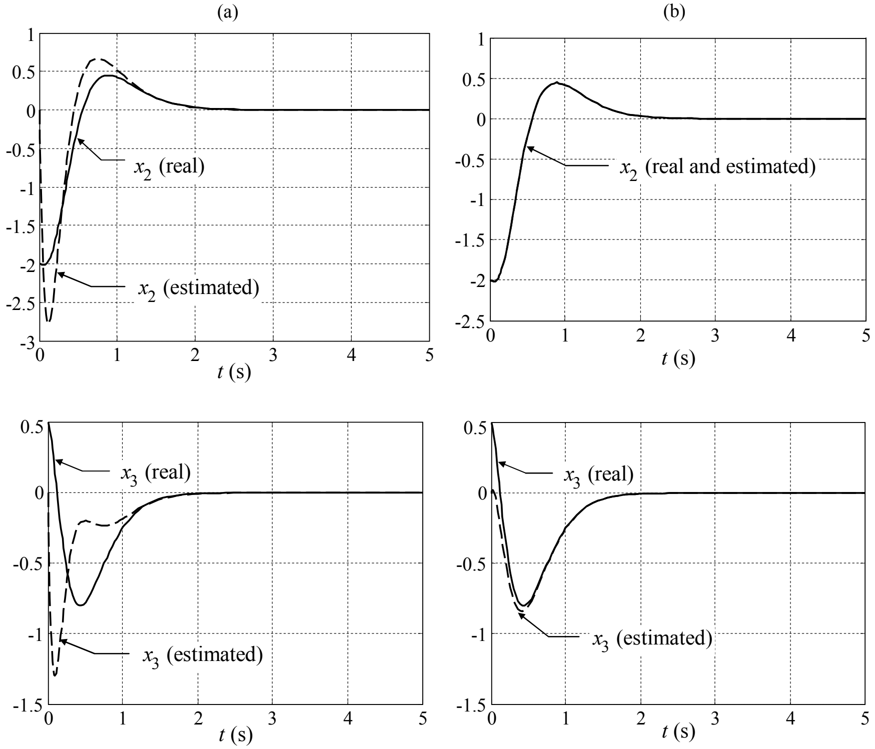


Fig. 2.11 Free responses of real and estimated states, in case of:
(a) a full-state observer (order 4); (b) a reduced-order observer (order 2).

The comparison of the state variable x_3 plots in the two sides of the figure shows that the reconstruction of the reduced-order observer converges faster to the real state than that of the *n*th-order observer. This result was expected, since the first observer has two poles less than the second at the same values $s=-6$. This constitutes an advantage of the reduced-order observer, its drawback being that it does not offer anymore the possibility of comparing estimated and measured values, which is at the base of the instrument failure detection methods by analytical redundancy.

The plots concerning the state variable x_2 in the case of the reduced-order observer are undistinguishable right from the initial time. This results from the fact that this state, just as x_1, is no longer reconstructed by this observer but simply transmitted from the measurement $y_2 = x_2$.

(e) Let us design now deliberately an identity observer much slower than the closed-loop system, so as to illustrate better its influence. With a choice of all its 4 poles located at $s=-2$, the design by the general formula, with the controllability canonical form, and arbitrary parameter vectors $\begin{bmatrix}0 & 1\end{bmatrix}^{\mathrm{T}}$, $\begin{bmatrix}1 & 0\end{bmatrix}^{\mathrm{T}}$, $\begin{bmatrix}0 & 0\end{bmatrix}^{\mathrm{T}}$ and $\begin{bmatrix}1 & -2\end{bmatrix}^{\mathrm{T}}$, yields the matrices:

$$\mathbf{G}=\begin{pmatrix} 0.380 & 0.409 \\ -3.530 & 5.569 \\ 0.050 & -0.140 \\ -13.284 & 29.568 \end{pmatrix}, \ \mathbf{F}=\mathbf{A}-\mathbf{G}\mathbf{C}_m=\begin{pmatrix} -0.380 & -0.409 & -1.950 & 0 \\ 3.530 & -5.569 & 0 & 1 \\ -0.050 & 0.011 & -1.915 & 0.001 \\ 13.284 & -8.095 & 26.339 & -0.136 \end{pmatrix},$$

with $\mathbf{J}=\mathbf{B}$, whereas $\mathbf{H}$ and $\mathbf{K}$ are given by (2.50). Fig. 2.12 shows the response of the system, closed through this observer, to a square wave reference of unit amplitude, from an initial plant state $\mathbf{x}_0 = \begin{pmatrix}0 & -2 & 0.5 & 0\end{pmatrix}^{\mathrm{T}}$ different from that of the observer $\mathbf{z}_0 = 0$.

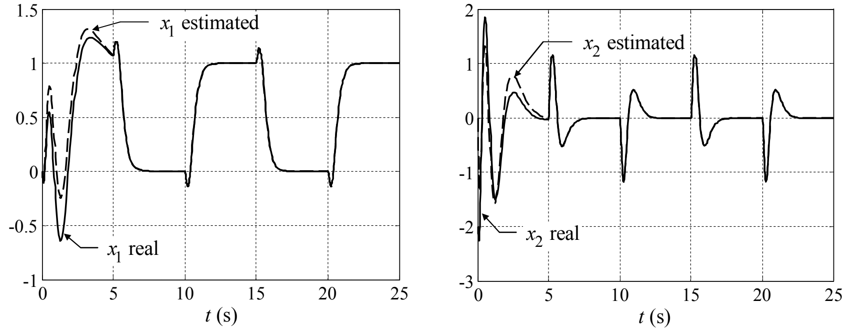


Fig. 2.12 Closed-loop square wave response for unequal plant and observer initial states.

This figure illustrates well the fact that, once the estimated state has converged to the real state, the observer dynamics does not appear anymore in the response to *reference* steps*:* the observer poles have no more effect. However, a new disturbance of the plant initial state would trigger a new transient response of the observer, with its own dynamics (see comments of Sect. 2.5.5).

Exercise 2.2 Dry Friction Compensation by a Disturbance Observer.

Consider again the inverted pendulum LIP 100 (Amira). The goal is here to compensate the dry friction of the cart on its rail by means of a disturbance observer. This study will be performed in discrete time.

a) Assume here that the plant has three measurements: x_c, θ and $\dot{x}_c$. Calculate first a control law for the discrete-time model of this plant, by placing all closed-loop continuous eigenvalues at $\lambda_{Li} = -5\,\mathrm{s}^{-1}$, $i = 1, \cdots, 4$.

b) Calculate a reduced-order observer, which allows on one hand to reconstruct the missing state variable, $x_4 = \dot{\theta}$, on the other hand to compensate the dry friction, which will be modeled as a constant disturbance, of unknown amplitude and changing sign at each direction change of the cart.

c) Simulate the closed loop by including this disturbance in the diagram, and evaluate the efficiency of the solution developed in question (b).

Solution:

(a) The plant is equipped here with the three-measurement matrix $\mathbf{C}_m = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$, which should be substituted in the program *MMCE.m* to the default measurement matrix by the statement change to three measurements.

The control law calculated by the pole placement method is obviously the same as the one calculated in question (d) of Exercise 1.1, since it does not depend on the number of state variables accessible to measurement:

$$\mathbf{L} = \boldsymbol{\ell}^{\mathrm{T}} = \begin{pmatrix} -2.0298 & 2.0162 & 3.5714 & 0.4434 \end{pmatrix}, \quad M = -2.0298\,.$$

(b) The disturbance observer will be calculated as in Sects. 2.6.2.3 and 2.6.2.4, where a functional observer will be replaced by a reduced-order observer.

The *MMCE.m* program allows performing these calculations by choosing in the Main menu, the synthesis of a disturbance observer, option available only with the inverted pendulum and in discrete time. The program does the following: first, the continuous-time disturbance-observer model is incorporated into the continuous-time plant model, as in Sect. 2.6.2.4; then the augmented system is sampled at the period $T_s = 0.03\,s$.

The continuous-time model augmented by the constant disturbance has, according to (2.78), the state representation:

$$\tilde{\mathbf{A}} = \begin{pmatrix} 0 & 0 & -1.950 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & -0.129 & -1.915 & 0.0008 & -6.134 \\ 0 & 21.473 & 26.339 & -0.1362 & 84.30 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},\ \tilde{\mathbf{B}} = \begin{pmatrix} 0 \\ 0 \\ -6.134 \\ 84.30 \\ 0 \end{pmatrix},\ \tilde{\mathbf{C}}^{\mathrm{T}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

This augmented model is then sampled with zero-order hold by the statement c2dm(A_tild,B_tild,C_tild,D_tild,Ts,'zoh'). The result is the discrete time state representation below, which is then partitioned as shown, since the state variables accessible to measurement are the first three ones:

$$\tilde{\mathbf{\Phi}} = \left(\begin{array}{ccc:cc} 1 & 0.0001 & -0.0569 & 4.1\times10^{-7} & 0.0053 \\ 0 & 1.0097 & 0.0116 & 0.0300 & 0.0372 \\ 0 & -0.0038 & 0.9442 & -3.3\times10^{-5} & -0.1789 \\ \hdashline 0 & 0.6434 & 0.7688 & 1.0056 & 2.4607 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right) \begin{array}{l} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} q \\ \left.\begin{matrix} \\ \\ \end{matrix}\right\} n-q \end{array},\ \tilde{\mathbf{\Gamma}} = \left(\begin{array}{c} 0.0053 \\ 0.0372 \\ -0.1789 \\ \hdashline 2.4607 \\ 0 \end{array}\right).$$

At the difference to Sect. 2.6.2.4, instead of estimating only a linear combination of part of the state variables and of the disturbance, as in (2.80), the reduced-order observer will have the duty to estimate the missing part of the augmented state vector $\begin{pmatrix}\mathbf{x}_k^{\mathrm{T}} & \xi_\kappa^{\mathrm{T}}\end{pmatrix}^{\mathrm{T}}$ of dimension 5. The control law, applied to that partially reconstructed state vector, will then be

$$u = -\begin{pmatrix}\boldsymbol{\ell}^{\mathrm{T}} & 1\end{pmatrix}\begin{pmatrix}\mathbf{x}_k \\ \xi_k\end{pmatrix} = -\begin{pmatrix}\ell_1 & \ell_2 & \ell_3 & \ell_4 & 1\end{pmatrix}\begin{pmatrix}\mathbf{x}_k \\ \xi_k\end{pmatrix}.$$

The design of this reduced-order observer, of order $n-q=5-3=2$, by pole placement, with two poles at $-10\ \mathrm{s}^{-1}$, yields the following matrices for its generalized state model:

$$\mathbf{G}_r = \begin{pmatrix} 0.3518 & 8.8015 & -11.9151 \\ 0.0428 & -0.0016 & -1.4481 \end{pmatrix},$$

$$\mathbf{F} = \begin{pmatrix} 0.7408 & 0 \\ 0 & 0.7408 \end{pmatrix}, \quad \mathbf{G} = \begin{pmatrix} -0.0912 & -1.7676 & 3.1094 \\ -0.0111 & -0.0050 & 0.2969 \end{pmatrix}, \quad \mathbf{J} = \begin{pmatrix} 0 \\ -0.2592 \end{pmatrix},$$

$$\mathbf{H} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{K} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0.3518 & 8.8015 & -11.9151 \\ 0.0428 & -0.0016 & -1.4481 \end{pmatrix}.$$

(c) The simulation is made by means of the third simulation diagram of Appendix D. In this diagram, the observer is broken down into the various matrices which compose its representation in the generalized observer form. A Coulomb & Viscous Friction block serves to model the dry friction, the viscous friction coefficient being set to zero there. On the real setup where the experiments have been performed, the dry friction represents a threshold of about 2.5 V at the point where the control voltage is applied to the plant.

The switch S_2 of the diagram is used to enable or disable this disturbance term, while the switches S_3 and S_4 allow choosing between compensation by disturbance observer, compensation by fixed threshold with sign switched at each cart travel direction change, and no compensation at all.

The two oscillograms in Fig. 2.13, recorded with a square wave reference signal of 2 volts amplitude and the same scales, point out the limit cycle resulting from the nonlinearity constituted by the dry friction. The substantial gain brought by the compensation by observer can be seen. It is clear that the fixed threshold compensation is perfect only if the intensity of the dry friction is known exactly, and if it does not evolve with time, situations rarely encountered in practice.

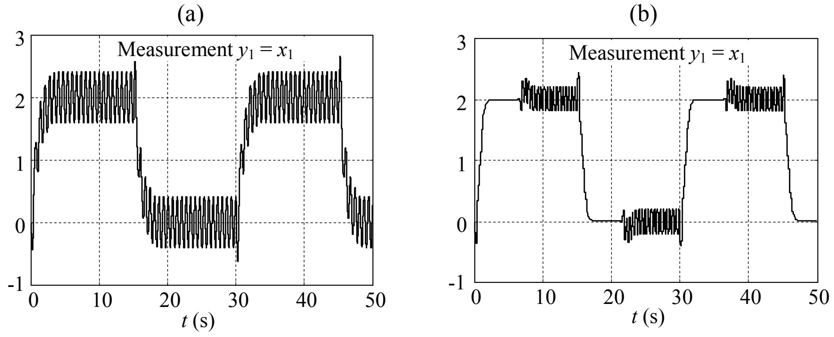


Fig. 2.13 Closed-loop square wave response of the cart position, (a) without dry friction disturbance compensation, (b) with disturbance observer.

3 Optimal Control

In this chapter and the following, a completely different approach will be taken than in the previous chapters: instead of considering the eigenvalues of the closed-loop control system or those of an observer, we will invoke quadratic optimization methods, deterministic here, stochastic in Chapter 4. For this reason, the dynamic behavior of the control system is no longer mastered directly. Fortunately however, the control designed in this chapter will have robustness properties, going thus beyond the simple stability requirement, as will be seen at the end of the chapter.

At the difference with other books using the *dynamic programming* approach and the Bellmann's *Principle of Optimality*, as in [AlSa04], the approach taken in the present book to establish the optimal control law will be based on the *calculus of variations* and the *Lagrangian multipliers*, which are the mathematical bases for many optimization problems in Physics and will allow us to make a parallel with the equations of motion of *Analytical Mechanics*. The correspondence with the dynamic programming will be discussed at the end of this chapter. The calculus of variations is also the method used in [Gee07], where closed-loop state-feedback optimal control laws are derived directly from the Bellman's principle in one step.

3.1 Introduction

In control theory, there are two different kinds of optimization:

- the *parametric* optimization, which consists in looking for the optimal *parameters* of a controller $C(s)$ of imposed structure, e.g. the proportional and integral coefficients K_p and T_i of a PI controller, by trying to minimize a *function J* of these parameters, as e.g. $J=\int_0^\infty \varepsilon^2(t)dt$; one tries then to find K_p and T_i such that $\dfrac{\partial J}{\partial K_p}=0$ and $\dfrac{\partial J}{\partial T_i}=0$;

E. Ostertag, *Mono- and Multivariable Control and Estimation,*
DOI 10.1007/978-3-642-13734-1_3, © Springer-Verlag Berlin Heidelberg 2011

- the *functional* optimization, where it is assumed on the contrary that the controller structure is completely free and thus that the control law $\mathbf{u}(t)$ or $\mathbf{u}_k$ which will be applied to the plant input can be chosen arbitrarily. To a given function $\mathbf{u}(t)$ or $\mathbf{u}_k$ is associated a given value of some criterion *J*, which is thus a *functional*. By language abuse, *J* is often called simply a cost *function*.

It is this second kind of optimization which will be the topic of this chapter. The objective is to find the *function* $u(t)$, respectively $\mathbf{u}(t)$ in the MIMO case, which minimizes the criterion *J*, which is always scalar and which depends on this function, hence the appellation *functional*, or *structural*, *optimization*.

Mathematical fundamentals. The variational calculus has been introduced by the Bernoulli brothers, who studied with this mathematical tool the well known brachistochrone problem, in 1697.

In the 18th century, Euler and Lagrange established the basics of variational calculus. The Euler-Lagrange equation was used later mainly in Physics, in the fields of Optics and Mechanics: Fermat's Principle for the propagation of light in media with different refringencies, Least Action Principle of Maupertuis (18th century) and of Hamilton for the determination of motions in Analytical Mechanics, in the 19th century. The work of Hamilton in Analytical Mechanics has been continued later by Jacobi.

The technical development of the 20^{th} century imposed the taking into account of constraints. Let us mention the works of Bellman (1950), of Feldbaum (1953) and of Pontryagin (1956), if only the most important milestones should be highlighted.

3.2 Free Final State Optimization

3.2.1 Problem Description

This is the basic problem of Optimal Control. The hypotheses are the following. Consider

- a non necessarily linear, nor time-invariant system:

$$\dot{\mathbf{x}}(t) = \mathbf{f}\left[\mathbf{x}(t), \mathbf{u}(t), t\right], \tag{3.1}$$

$$\mathbf{x}(0) = \mathbf{x}_0 \text{ given,}$$

$$\mathbf{u}(t) \in \mathscr{U} \; ;$$

- and a cost functional:

$$J = h\left[\mathbf{x}(t_f)\right] + \int_0^{t_f} f_0\left[\mathbf{x}(t),\mathbf{u}(t),t\right]dt\,, \tag{3.2}$$

where t_f is the final time, which is fixed here, and *J*, *h* and f_0 are scalars.

The objective is to find the control law $\mathbf{u}(t)$, in the universe $\mathscr{U}$ of all possible control laws, which minimizes *J*. The choice of this cost functional results as a matter of fact from the combination of several criteria [FöRo88].

The integral term expresses the desire to minimize, all along the trajectory, the energy consumed by the plant, at the input of which the control law $\mathbf{u}(t)$ will be applied, the unnecessary deviations from its state space trajectory between the initial state $\mathbf{x}_0$ and the expected final state $\mathbf{x}(t_f)$, deviations which affect of course the variations of the output $\mathbf{y}(t) = \mathbf{C}\,\mathbf{x}(t)$ on this path , and the time t_f itself.

Very often, as will be seen in more details in Sect. 3.5, the used criteria are of quadratic type, so as to take into account variations of $\mathbf{u}(t)$ or $\mathbf{x}(t)$ of opposite signs equally. Therefore, energy and optimal trajectory criteria will often enter into the previous f_0 function as two quadratic forms, yielding thus a cost functional having the following form:

$$J = \frac{1}{2}\int_0^{t_f} (\mathbf{x}^{\mathrm{T}}\mathbf{Q}\,\mathbf{x} + \mathbf{u}^{\mathrm{T}}\mathbf{R}\,\mathbf{u})dt\,.$$

The factor $1/2$ is not important and facilitates only the subsequent calculations.

A minimization of the trajectory duration, thus of t_f, can also be written in the form of an integral criterion:

$$J = \int_0^{t_f} 1\cdot dt\,.$$

The two previous criteria, quadratic criterion and time optimality, are grouped in the more general form, called Lagrangian cost functional:

$$J = \int_0^{t_f} f_0\left[\mathbf{x}(t),\mathbf{u}(t),t\right]dt\,,$$

which represents the second term of the cost functional (3.2), and which will be encountered again in the case of optimization problems with entirely fixed final state, in Sect. 3.3.

It remains to justify the first term of expression (3.2). At the difference to the previous one, this criterion deals only with the final state. It intervenes in problems where the final state is crucial, such as the hit of a projectile on a moving target or the accuracy of moon landing, as far as the height coordinate of the Lunar Module above the landing point is concerned. In its most general form, namely

$$J = h\left[\mathbf{x}(t_f), t_f\right],$$

this criterion is called Mayer criterion, from the name of a mathematician of the 19th century end.

The cost functional resulting from the grouping of the two previous criteria is called cost functional of Bolza, a mathematician of the beginning of the 20th century, and constitutes very closely the cost functional (3.2), which will be used in this whole chapter.

3.2.2 Modified Cost Functional. Hamiltonian

The problem encountered here is typical of *variational calculus*. In order to characterize an optimal control, one tries to evaluate the effect of a small variation of $\mathbf{u}(t)$ on the cost functional and to impose that it does not produce any diminution of the last. The approach will thus have three steps:

1. start from a given control law $\mathbf{u}(t)$, assumed optimal;
2. let $\mathbf{u}(t)$ vary slightly and calculate the corresponding variation of the cost functional *J*;
3. this variation should be positive (non decreasing) if the starting control law $\mathbf{u}(t)$ was optimal, thus if *J* was minimal.

3.2.2.1 Method of Lagrange Multipliers

Due to the existence of an *auxiliary condition*, namely the differential equation (3.1), which can be written as

$$\dot{\mathbf{x}}(t) - \mathbf{f}\left[\mathbf{x}(t), \mathbf{u}(t), t\right] = 0,$$

and which causes any variation of $\mathbf{u}(t)$ to result in a variation of $\mathbf{x}(t)$, it is not possible to solve this problem in a direct way.

The problem is then solved indirectly by the *method of Lagrange multipliers*, which allows transforming an extremum problem *with* auxiliary condition to an extremum problem *without* auxiliary condition.

The multiplier considered here is a vector of same dimension as $\mathbf{x}$ and depending on the time: $\boldsymbol{\psi}(t) = \left(\psi_1(t) \quad \cdots \quad \psi_n(t)\right)^{\mathrm{T}}$.

The method consists in adding to the integrand a *permanently vanishing* term, which contains the Lagrange multipliers. The following *modified cost functional*,

$$J = h\left[\mathbf{x}(t_f)\right] + \int_0^{t_f} f_0\left[\mathbf{x}(t), \mathbf{u}(t), t\right] dt + \underbrace{\int_0^{t_f} \boldsymbol{\psi}^{\mathrm{T}}(t)\left\{\dot{\mathbf{x}}(t) - \mathbf{f}\left[\mathbf{x}(t), \mathbf{u}(t), t\right]\right\} dt}_{\equiv 0}, \quad (3.3)$$

is thus built, which is identical to the previous one. Note that the auxiliary condition is now *included in J*. It is thus no more necessary to account for it separately.

3.2.2.2 Hamiltonian

In order to shorten the equations, one introduces the *Hamiltonian:*

$$H(\boldsymbol{\psi}, \mathbf{x}, \mathbf{u}, t) = -f_0(\mathbf{x}, \mathbf{u}, t) + \boldsymbol{\psi}^{\mathrm{T}} \cdot \mathbf{f}(\mathbf{x}, \mathbf{u}, t) \ . \quad (3.4)$$

Remark 3.1. H is a scalar function of $\mathbf{u}(t)$, of $\mathbf{x}(t)$, of a presently unknown vector $\boldsymbol{\psi}(t)$, which will be defined later, and also, in the case for instance of time-varying systems, an explicit function of time.

By substituting (3.4) into (3.3), we get:

$$J = h\left[\mathbf{x}(t_f)\right] + \int_0^{t_f} \left\{\boldsymbol{\psi}^{\mathrm{T}}(t)\, \dot{\mathbf{x}}(t) - H\left[\boldsymbol{\psi}(t), \mathbf{x}(t), \mathbf{u}(t), t\right]\right\} dt. \quad (3.5)$$

Remark 3.2. For time-invariant systems, the explicit time dependence disappears from the state equation: $\dot{\mathbf{x}}(t) = \mathbf{f}\left[\mathbf{x}(t), \mathbf{u}(t)\right]$. If, in addition, the integrand f_0 of the minimization criterion does not depend explicitly on time, the explicit time dependence disappears also from the two following equations:

$$f_0 = f_0\left[\mathbf{x}(t), \mathbf{u}(t)\right],$$

$$H = H\left[\boldsymbol{\psi}(t), \mathbf{x}(t), \mathbf{u}(t)\right].$$

This brings about a simplification of the writings, which will be used in the rest of this chapter. For the same purpose, the time dependencies of the functions which are arguments of the above expressions will also be omitted, as long as no ambiguity results: $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u})$, $f_0 = f_0(\mathbf{x}, \mathbf{u})$, $H = H(\boldsymbol{\psi}, \mathbf{x}, \mathbf{u})$.

3.2.3 Variational Calculus

As a starting point, suppose that a given nominal control law $\mathbf{u}(t)$, which satisfies the constraints $\mathbf{u}(t) \in \mathcal{U}$, has been found. This control $\mathbf{u}(t)$ produces some corresponding state trajectory, $\mathbf{x}(t)$.

To this control $\mathbf{u}(t)$ a *small* variation, or perturbation, $\delta\mathbf{u}(t)$, is then applied, which transforms from that time on the initial control law to $\mathbf{u}(t)+\delta\mathbf{u}(t)$. This infinitesimally small variation is such that:

$$\int_0^{t_f} \left|\delta u_i(t)\right| dt < \varepsilon \,, \tag{3.6}$$

whatever ε, arbitrarily small, and this for all components of **u**, thus for every *i*.

The perturbation on $\mathbf{u}(t)$ causes obviously a small variation (perturbation) of the system state trajectory, which becomes from that time on $\mathbf{x}(t)+\delta\mathbf{x}(t)$, as illustrated in Fig. 3.1 in the case of a one-dimensional problem.

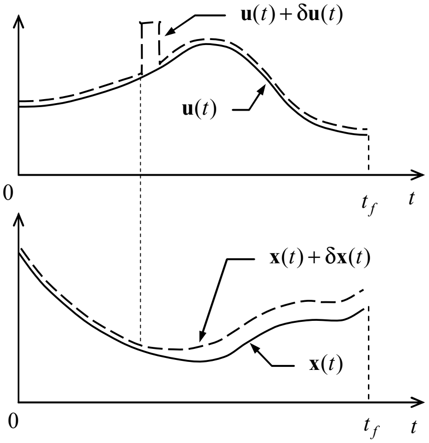


Fig. 3.1 Perturbed control law and resulting trajectory.

The trajectory perturbation, $\delta\mathbf{x}(t)$, is itself an infinitesimally small quantity of first order in ε for every *t*, as can be seen by applying (3.6) since the state depends essentially on the integral of the control law.

The corresponding cost functional variation amounts to

$$\delta J = h\left[\mathbf{x}(t_f)+\delta\mathbf{x}(t_f)\right] - h\left[\mathbf{x}(t_f)\right] \\ + \int_0^{t_f} \left\{ \boldsymbol{\psi}^{\mathrm{T}}\,\delta\dot{\mathbf{x}} - H\left[\boldsymbol{\psi}, \mathbf{x}+\delta\mathbf{x}, \mathbf{u}+\delta\mathbf{u}\right] + H\left[\boldsymbol{\psi},\mathbf{x},\mathbf{u}\right] \right\} dt \,. \tag{3.7}$$

The first term of the integral is evaluated by integration by parts:

$$\int_0^{t_f} \boldsymbol{\psi}^{\mathrm{T}}(t)\,\delta\dot{\mathbf{x}}(t)\,dt = \boldsymbol{\psi}^{\mathrm{T}}(t)\,\delta\mathbf{x}(t)\Big|_0^{t_f} - \int_0^{t_f} \dot{\boldsymbol{\psi}}^{\mathrm{T}}(t)\,\delta\mathbf{x}(t)\,dt$$

$$= \boldsymbol{\psi}^{\mathrm{T}}(t_f)\,\delta\mathbf{x}(t_f) - \boldsymbol{\psi}^{\mathrm{T}}(0)\,\delta\mathbf{x}(0) - \int_0^{t_f} \dot{\boldsymbol{\psi}}^{\mathrm{T}}\,\delta\mathbf{x}\,dt\,. \qquad (3.8)$$

To calculate the second term of the integral, we will use its Taylor series expansion and its first order approximation, i.e. at the order of the infinitesimally small ε introduced above. Let us first visualize separately the two variations, by subtracting from the integrand, and adding to it, the term $H(\boldsymbol{\psi},\mathbf{x},\mathbf{u}+\delta\mathbf{u})$:

$$\int_0^{t_f} \left[H(\boldsymbol{\psi},\mathbf{x}+\delta\mathbf{x},\mathbf{u}+\delta\mathbf{u}) - H(\boldsymbol{\psi},\mathbf{x},\mathbf{u})\right] dt =$$

$$\int_0^{t_f} \left[H(\boldsymbol{\psi},\mathbf{x}+\delta\mathbf{x},\mathbf{u}+\delta\mathbf{u}) - H(\boldsymbol{\psi},\mathbf{x},\mathbf{u}+\delta\mathbf{u}) + H(\boldsymbol{\psi},\mathbf{x},\mathbf{u}+\delta\mathbf{u}) - H(\boldsymbol{\psi},\mathbf{x},\mathbf{u})\right] dt\,. \quad (3.9)$$

Apply the first order Taylor series expansion approximation to the first two terms of the integrand. The following approximation holds for the previous integral:

$$\simeq \int_0^{t_f} \left[\frac{\partial}{\partial \mathbf{x}^{\mathrm{T}}} H(\boldsymbol{\psi},\mathbf{x},\mathbf{u}+\delta\mathbf{u})\,\delta\mathbf{x} + H(\boldsymbol{\psi},\mathbf{x},\mathbf{u}+\delta\mathbf{u}) - H(\boldsymbol{\psi},\mathbf{x},\mathbf{u})\right] dt\,.$$

Subtract now from the integrand, and add immediately again to it, the term $\left[\partial H(\boldsymbol{\psi},\mathbf{x},\mathbf{u})/\partial \mathbf{x}^{\mathrm{T}}\right]\delta\mathbf{x}$. The above integral reads thus:

$$= \int_0^{t_f} \left\{\left[\frac{\partial}{\partial \mathbf{x}^{\mathrm{T}}} H(\boldsymbol{\psi},\mathbf{x},\mathbf{u}+\delta\mathbf{u}) - \frac{\partial}{\partial \mathbf{x}^{\mathrm{T}}} H(\boldsymbol{\psi},\mathbf{x},\mathbf{u})\right]\delta\mathbf{x}\right.$$

$$\left. + \frac{\partial}{\partial \mathbf{x}^{\mathrm{T}}} H(\boldsymbol{\psi},\mathbf{x},\mathbf{u})\,\delta\mathbf{x} + H(\boldsymbol{\psi},\mathbf{x},\mathbf{u}+\delta\mathbf{u}) - H(\boldsymbol{\psi},\mathbf{x},\mathbf{u})\right\} dt\,,$$

or:

$$= \left\{\int_0^{t_f} \left[\frac{\partial}{\partial \mathbf{x}^{\mathrm{T}}} H(\boldsymbol{\psi},\mathbf{x},\mathbf{u}+\delta\mathbf{u}) - \frac{\partial}{\partial \mathbf{x}^{\mathrm{T}}} H(\boldsymbol{\psi},\mathbf{x},\mathbf{u})\right] dt\right\}\delta\mathbf{x}$$

$$+ \int_0^{t_f} \left[\frac{\partial}{\partial \mathbf{x}^{\mathrm{T}}} H(\boldsymbol{\psi},\mathbf{x},\mathbf{u})\,\delta\mathbf{x} + H(\boldsymbol{\psi},\mathbf{x},\mathbf{u}+\delta\mathbf{u}) - H(\boldsymbol{\psi},\mathbf{x},\mathbf{u})\right] dt\,. \quad (3.10)$$

The first of these two integrals is an infinitesimally small quantity of first order in ε, for the reasons invoked above, concerning the effect of the control perturbation on the state trajectory, by taking into account in addition to the hypothesis

(3.6) the expression (3.4) of the Hamiltonian. Since this integral is multiplied to the right by $\delta\mathbf{x}$, itself infinitesimally small of first order in ε, as was proved above, the product of these two factors is thus infinitesimally small of the order of ε^2, which makes it negligible compared to the second integral of (3.10).

We obtain thus for the expression (3.9) the following approximation:

$$\int_0^{t_f}\left[H(\psi,\mathbf{x}+\delta\mathbf{x},\mathbf{u}+\delta\mathbf{u})-H(\psi,\mathbf{x},\mathbf{u})\right]dt$$
$$\simeq\int_0^{t_f}\left[\frac{\partial}{\partial\mathbf{x}^{\mathrm{T}}}H(\psi,\mathbf{x},\mathbf{u})\,\delta\mathbf{x}+H(\psi,\mathbf{x},\mathbf{u}+\delta\mathbf{u})-H(\psi,\mathbf{x},\mathbf{u})\right]dt\,. \tag{3.11}$$

Substituting (3.8) and (3.11) into (3.7), the following expression is finally obtained for the cost functional variation:

$$\delta J\simeq\left.\frac{dh}{d\mathbf{x}^{\mathrm{T}}}\right|_{t=t_f}\delta\mathbf{x}(t_f)+\psi^{\mathrm{T}}(t_f)\delta\mathbf{x}(t_f)-\psi^{\mathrm{T}}(0)\delta\mathbf{x}(0)-\int_0^{t_f}\dot{\psi}^{\mathrm{T}}\,\delta\mathbf{x}\,dt$$
$$-\int_0^{t_f}\left[\frac{\partial}{\partial\mathbf{x}^{\mathrm{T}}}H(\psi,\mathbf{x},\mathbf{u})\,\delta\mathbf{x}+H(\psi,\mathbf{x},\mathbf{u}+\delta\mathbf{u})-H(\psi,\mathbf{x},\mathbf{u})\right]dt\,.$$

After some rearrangement of the terms and the use of (B.7) of Appendix B, this expression is also written:

$$\delta J\simeq\left[\left.\frac{dh}{d\mathbf{x}}\right|_{t=t_f}+\psi(t_f)\right]^{\mathrm{T}}\delta\mathbf{x}(t_f)-\psi^{\mathrm{T}}(0)\delta\mathbf{x}(0)-\int_0^{t_f}\left[\dot{\psi}+\frac{\partial}{\partial\mathbf{x}}H(\psi,\mathbf{x},\mathbf{u})\right]^{\mathrm{T}}\delta\mathbf{x}\,dt$$
$$-\int_0^{t_f}\left[H(\psi,\mathbf{x},\mathbf{u}+\delta\mathbf{u})-H(\psi,\mathbf{x},\mathbf{u})\right]dt\,. \tag{3.12}$$

Note first that $\delta\mathbf{x}(0)=0$, since a control variation does not modify the initial state. The second term of the right side of (3.12) disappears thus.

The goal of the present calculation being to find an extremum of our cost functional, we want to realize $\delta J=0$.

3.2.4 Adjoint Equation

Since the vector $\psi(t)$ is still arbitrary at this state of the calculation, let us choose it so as to cancel the first two terms on the right side of (3.12). It is enough for this to choose $\psi(t)$ as the solution of the *adjoint differential equation*

$$\dot{\boldsymbol{\psi}}(t) = -\frac{\partial}{\partial \mathbf{x}} H(\boldsymbol{\psi}, \mathbf{x}, \mathbf{u}) \ , \tag{3.13}$$

which satisfies the *final condition* (also called *transversality condition*):

$$\boldsymbol{\psi}(t_f) = -\left.\frac{dh}{d\mathbf{x}}\right|_{t=t_f} \ . \tag{3.14}$$

Remark 3.3. With (3.4) and (B.16) the differential equation (3.13) becomes

$$-\dot{\boldsymbol{\psi}}(t) = -\frac{\partial f_0}{\partial \mathbf{x}} + \frac{\partial \mathbf{f}^{\mathrm{T}}}{\partial \mathbf{x}} \boldsymbol{\psi}(t) \ . \tag{3.15}$$

This relation constitutes a system of n linear differential equations which, with the final conditions (3.14), allows determining the n components of $\boldsymbol{\psi}(t)$. Note that, due to the presence of a minus sign at the left side, their numerical solving will occur *backwards*, i.e. at *decreasing* time, starting from the *final conditions*.

Remark 3.4. The vector $\boldsymbol{\psi}(t)$ is called *costate vector*.

3.2.5 Hamilton's Equations

The Hamiltonian definition (3.4) and the formula of differentiation (B.8) show readily that

$$\frac{\partial H}{\partial \boldsymbol{\psi}} = \frac{\partial}{\partial \boldsymbol{\psi}} (\boldsymbol{\psi}^{\mathrm{T}} \cdot \mathbf{f}) = \mathbf{f} \ .$$

It follows, according to (3.1), that

$$\frac{d\mathbf{x}}{dt} = \dot{\mathbf{x}} = \frac{\partial H}{\partial \boldsymbol{\psi}} \ . \tag{3.16}$$

This equation, together with (3.13) repeated here:

$$\frac{d\boldsymbol{\psi}}{dt} = \dot{\boldsymbol{\psi}} = -\frac{\partial H}{\partial \mathbf{x}} \ , \tag{3.17}$$

represent the *Hamilton's (conjugated) equations*, well known in Analytical Mechanics.

3.2.6 Similarity between Analytical Mechanics and Optimal Control

Obviously a full similarity exists between the extremum problem tackled here and the *Hamilton's Principle* of the Analytical Mechanics, which states that, among all motions transferring a given system from a point M_0 of its configuration space at initial time t_0 to a point M_1 at time t_1, the real motion is the one for which the *action integral* $I = \int_{t_0}^{t_1} L\,dt$ is extremum, i.e. $\delta I = 0$, where L is the system's *Lagrangian*. Table 3.1 illustrates this formal similarity.

Table 3.1 Formal similarity between Analytical Mechanics and Optimal Control.

Analytical Mechanics	Optimal Control
Generalized coordinates of n mass points:[1] $\mathbf{q} = \{q_j\},\ j = 1,\dots,f$	System state variables and state vector: $\mathbf{x} = \{x_i\},\ i = 1,\dots,n$
Lagrangian: L	Function involved in the cost criterion J: $f_0(\mathbf{x},\mathbf{u},t)$
Integral (criterion) to minimize: $I = \int_{t_0}^{t_1} L\,dt$	Cost functional (for fixed final state): $J = \int_0^{t_f} f_0\left[\mathbf{x}(t),\mathbf{u}(t),t\right]dt$
System's Hamiltonian: $H = -L + \sum_{j=1}^{f} p_j\,\dot{q}_j$	System's Hamiltonian: $H = -f_0(\mathbf{x},\mathbf{u},t) + \boldsymbol{\psi}^{\mathrm{T}}(t)\cdot\mathbf{f}(\mathbf{x},\mathbf{u},t)$ $= -f_0(\mathbf{x},\mathbf{u},t) + \sum_{i=1}^{n} \psi_i\,\dot{x}_i$
Canonical conjugated momentum of q_j: $\mathbf{p} = \{p_j\},\ j = 1,\dots,f$	Costate vector: $\boldsymbol{\psi} = \{\psi_i\},\ i = 1,\dots,n$
Conjugated Hamilton's equations: $\begin{cases} \dot{q}_j = \dfrac{\partial H}{\partial p_j} \\ \dot{p}_j = -\dfrac{\partial H}{\partial q_j} \end{cases}$	Conjugated Hamilton's equations: $\begin{cases} \dot{\mathbf{x}} = \dfrac{\partial H}{\partial \boldsymbol{\psi}} \\ \dot{\boldsymbol{\psi}} = -\dfrac{\partial H}{\partial \mathbf{x}} \end{cases}$

[1].The number f of generalized coordinates is the number of degrees of freedom of the set of n mass points, i.e. the number of independent coordinates remaining once the constraints (links between points) have been taken into account

3.2.7 Pontryagin's Maximum Principle (1956)

Assuming that the hypotheses introduced by (3.13) and (3.14) are satisfied, the cost functional variation δJ, given by (3.12), becomes:

$$\delta J \simeq -\int_0^{t_f} \left[H(\boldsymbol{\psi}, \mathbf{x}, \mathbf{u}+\delta\mathbf{u}) - H(\boldsymbol{\psi}, \mathbf{x}, \mathbf{u})\right] dt\,. \tag{3.18}$$

The condition of optimality is then the following: for the optimal control law **u**, the Hamiltonian H is *maximum*, in other words

$$H\left[\boldsymbol{\psi}, \mathbf{x}, \mathbf{u}+\delta\mathbf{u}\right] \leq H\left[\boldsymbol{\psi}, \mathbf{x}, \mathbf{u}\right], \tag{3.19}$$

$\forall t \in \left[0, t_f\right]$ and $\forall\ (\mathbf{u}+\delta\mathbf{u}) \in \mathcal{U}$.

Proof by Contradiction. Suppose that, for some value of *t*, there exists a perturbed control law, $(\mathbf{u}+\delta\mathbf{u}) \in \mathcal{U}$, such that

$$H\left[\boldsymbol{\psi}(t), \mathbf{x}(t), \mathbf{u}(t)+\delta\mathbf{u}(t)\right] > H\left[\boldsymbol{\psi}(t), \mathbf{x}(t), \mathbf{u}(t)\right].$$

One could then modify the function **u**, as indicated in Fig. 3.1, in such a way that the integrand of (3.18) be positive in a small interval containing this value of *t*. This would result in a $\delta J < 0$, which would contradict the fact that the function **u** produces the minimum value of the cost functional *J*.

In conclusion, for every *t*, the particular value $\mathbf{u}(t)$ which corresponds to the optimal control law has the property to maximize the Hamiltonian. This result constitutes the *Pontryagin's Maximum Principle* for this problem.

3.2.8 Recapitulation: Theorem (Maximum's Principle)

Assume that $\mathbf{u}(t) \in \mathcal{U}$ *and* $\mathbf{x}(t)$ *represent respectively the optimal control law and the corresponding state trajectory, solutions of the optimal control problem defined at the beginning of this section. There exists then an adjoint trajectory* $\boldsymbol{\psi}(t)$ *such that the following holds, simultaneously:*

$$\begin{cases} \dot{\mathbf{x}}(t) = \mathbf{f}\left[\mathbf{x}(t), \mathbf{u}(t)\right] & \text{(plant state equation)} \\ \mathbf{x}(0) = \mathbf{x}_0 & \text{(plant initial state)} \end{cases},$$

$$\begin{cases} -\dot{\psi}(t) = -\dfrac{\partial f_0}{\partial \mathbf{x}} + \dfrac{\partial \mathbf{f}^{\mathrm{T}}}{\partial \mathbf{x}} \cdot \psi(t) & \text{(adjoint state equation)} \\ \psi(t_f) = -\dfrac{dh\left[\mathbf{x}(t_f)\right]}{d\mathbf{x}} & \text{(transversality condition)} \end{cases},$$

$\forall t \in \left[0, t_f\right]$ *and* $\forall\ (\mathbf{u} + \delta\mathbf{u}) \in \mathscr{U}$:

$$H\left[\psi(t), \mathbf{x}(t), \mathbf{u}(t) + \delta\mathbf{u}(t)\right] \leq H\left[\psi(t), \mathbf{x}(t), \mathbf{u}(t)\right].$$

This inequality is the maximum's condition, where H is the Hamiltonian:

$$H(\psi, \mathbf{x}, \mathbf{u}) = -f_0(\mathbf{x}, \mathbf{u}) + \psi^{\mathrm{T}} \cdot \mathbf{f}(\mathbf{x}, \mathbf{u}).$$

Remark 3.5. In some books, such as [Gee07], this result is given as the Pontryagin's *minimum* principle. This is simply due to the fact that these authors introduce a Hamiltonian with opposite sign as compared with our definition, their function f_0 and their costate vector ψ having opposite signs to ours. The x value which maximizes a function $y(x)$ is obviously the same which minimizes $-y(x)$.

3.3 Problems with Final State Constraints

In Sect. 3.2 we have assumed that the final state was entirely arbitrary: the tackled situation was called therefore *free final state* problem.

In many cases, however, the final state is confined to some domain of the state space. In this case, the previous problem hypotheses will be completed by a set of constraints on the final state:

$$x_i(t_f) = \tilde{x}_i, \quad i = 1, \ldots, r, \tag{3.20}$$

where the plant state variables have been renumbered if needed so as to place at the beginning the r ones which are submitted to constraints.

The search for conditions which must be satisfied by an optimal solution to this problem will occur by following the same approach as in Sect. 3.2. The modified cost functional J is created again, and the variation induced on J by a change of the control law from $\mathbf{u}(t)$ to a perturbed law $\mathbf{u}(t) + \delta\mathbf{u}(t)$ which satisfies also all the constraints, is evaluated.

Let us repeat here this variation, which has been established in (3.12):

$$\delta J \simeq \left[\left.\frac{dh}{d\mathbf{x}}\right|_{t=t_f} + \boldsymbol{\psi}(t_f) \right]^{\mathrm{T}} \delta\mathbf{x}(t_f) - \boldsymbol{\psi}^{\mathrm{T}}(0)\delta\mathbf{x}(0) - \int_0^{t_f} \left[\dot{\boldsymbol{\psi}} + \frac{\partial}{\partial \mathbf{x}} H(\boldsymbol{\psi},\mathbf{x},\mathbf{u}) \right]^{\mathrm{T}} \delta\mathbf{x}\, dt$$
$$- \int_0^{t_f} \left[H(\boldsymbol{\psi},\mathbf{x},\mathbf{u}+\delta\mathbf{u}) - H(\boldsymbol{\psi},\mathbf{x},\mathbf{u}) \right] dt .$$

Here again, $\delta\mathbf{x}(0) = 0$, and again we will choose $\boldsymbol{\psi}(t)$ so as to cancel as many terms as possible on the right side of this expression. Therefore we will impose again, as in (3.13),

$$\dot{\boldsymbol{\psi}}(t) = -\frac{\partial}{\partial \mathbf{x}} H(\boldsymbol{\psi},\mathbf{x},\mathbf{u}) ,$$

but the specification of the boundary conditions will be somewhat different from (3.14).

Indeed, since $x_i(t_f)$ is fixed for $i = 1,\ldots,r$, it ensues that $\delta x_i(t_f) = 0$ for these same values of index *i* for all the trajectories involved. It is thus not necessary to impose the condition (3.14) to *all* the components of $\boldsymbol{\psi}(t_f)$ to cancel the first term of (3.12). It will be enough to impose

$$\psi_i(t) = -\left.\left(\frac{dh}{d\mathbf{x}}\right)_i\right|_{t=t_f} = -\left.\frac{\partial h}{\partial x_i}\right|_{t=t_f} , \quad i = r+1,\ldots,n , \tag{3.21}$$

to ensure that the dot product of the first term of δJ vanishes, i.e. to obtain

$$\left[\left.\frac{dh}{d\mathbf{x}}\right|_{t=t_f} + \boldsymbol{\psi}(t_f) \right]^{\mathrm{T}} \delta\mathbf{x}(t_f) = 0 .$$

To summarize, the rule is the following:

- if $x_i(t_f)$ is submitted to constraint, $\psi_i(t_f)$ is free;
- if $x_i(t_f)$ is free, $\psi_i(t_f)$ is submitted to constraint.

Remark 3.6. If $\mathbf{x}(t_f)$ is entirely fixed, $\boldsymbol{\psi}(t_f)$ is entirely free. The transversality condition does therefore not apply anymore, and the term weighting the final state in the cost functional, namely $h\left[\mathbf{x}(t_f)\right]$, which is then constant, has no reason to stay there and can be suppressed from the expression (3.2) of *J*, which reduces thus to:

$$J = \int_0^{t_f} f_0 \left[\mathbf{x}(t), \mathbf{u}(t), t\right] dt . \tag{3.22}$$

3.4 Problems with Free Final Time

In some problems, the final time t_f is not fixed and can be chosen itself so as to yield the smallest possible cost functional. As an example, let us mention the trajectory of a moon rocket which minimizes the fuel consumption: there is no reason in this case to specify the landing instant.

If the final time t_f is left free, the optimization problem consists now in finding simultaneously $t_f > 0$ and $\mathbf{u}(t),\ 0 \le t \le t_f$, such that the cost functional J given by (3.2) is minimal.

It is clear that, if the best t_f value were known, one could impose it and would be left with the previous problem. All the conditions listed in the maximum's principle theorem must thus hold here in the same way. Only an additional condition is required here, from which the unknown value of t_f will be derived.

Consider as previously a modification of the control law $\mathbf{u}(t)$ to a new law, $\mathbf{u}(t) + \delta\mathbf{u}(t)$, which will produce a new trajectory $\mathbf{x}(t) + \delta\mathbf{x}(t)$, but this time with a new final time $t_f + dt_f$.

The new aspect of this problem is that the system final state variation is no longer $\delta\mathbf{x}(t_f)$, since the final time itself varies. The new final state is in reality $\mathbf{x}(t_f) + \delta\mathbf{x}(t_f + dt_f)$, and, if we denote it with $\mathbf{x}(t_f) + d\mathbf{x}(t_f)$, the following holds at first order approximation:

$$d\mathbf{x}(t_f) \simeq \delta\mathbf{x}(t_f) + \dot{\mathbf{x}}(t_f)\,dt_f = \delta\mathbf{x}(t_f) + \mathbf{f}\left[\mathbf{x}(t_f), \mathbf{u}(t_f)\right] dt_f . \tag{3.23}$$

A calculation similar to the one performed in Sect. 3.2 would lead to the following expression for the cost functional variation:

$$\delta J \simeq \left[\left.\frac{dh}{d\mathbf{x}}\right|_{t=t_f}\right]^{\mathrm{T}} d\mathbf{x}(t_f) + \boldsymbol{\psi}^{\mathrm{T}}(t_f)\,\delta\mathbf{x}(t_f) - \boldsymbol{\psi}^{\mathrm{T}}(0)\,\delta\mathbf{x}(0) + \underline{f_0\left[\mathbf{x}(t_f), \mathbf{u}(t_f)\right] dt_f}$$
$$- \int_0^{t_f} \left[\dot{\boldsymbol{\psi}} + \frac{\partial}{\partial \mathbf{x}} H(\boldsymbol{\psi}, \mathbf{x}, \mathbf{u})\right]^{\mathrm{T}} \delta\mathbf{x}\, dt - \int_0^{t_f} \left[H(\boldsymbol{\psi}, \mathbf{x}, \mathbf{u} + \delta\mathbf{u}) - H(\boldsymbol{\psi}, \mathbf{x}, \mathbf{u})\right] dt ,$$

which replaces (3.12).

The additional term $f_0\left[\mathbf{x}(t_f),\mathbf{u}(t_f)\right]dt_f$, which is underlined in the new expression, stems from the fact that, in the first order Taylor series expansion, the variation dt_f is also taken into account this time. Now, the explicit dependence of J on t_f is obtained from its partial derivative $\partial J/\partial t_f$, which according to (3.3) amounts to

$$\frac{\partial}{\partial t_f}\left[\int_0^{t_f}\left\{f_0\,(\mathbf{x},\mathbf{u})+\boldsymbol{\psi}^{\mathrm{T}}\left[\dot{\mathbf{x}}-\mathbf{f}\,(\mathbf{x},\mathbf{u})\right]\right\}dt\right]=f_0\left[\mathbf{x}(t_f),\mathbf{u}(t_f)\right],$$

where the formula for differentiation of a definite integral, whose limits are functions of the differential variable, (Leibnitz integral rule) has been used:

$$\frac{d}{dx}\left[\int_{a(x)}^{b(x)}\varphi(x,y)\,dy\right]=\int_{a(x)}^{b(x)}\frac{\partial\varphi}{\partial x}dy+\frac{db}{dx}\,\varphi\left[x,b(x)\right]-\frac{da}{dx}\,\varphi\left[x,a(x)\right].$$

By choosing again $\boldsymbol{\psi}(t)$ so as to satisfy the adjoint equation (3.13), by noting again also that $\delta\mathbf{x}(0)=0$, and by making the substitution

$$\delta\mathbf{x}(t_f)=d\mathbf{x}(t_f)-\mathbf{f}\left[\mathbf{x}(t_f),\mathbf{u}(t_f)\right]dt_f$$

derived from (3.23), we obtain:

$$\delta J\simeq\left[\left.\frac{dh}{d\mathbf{x}}\right|_{t=t_f}+\boldsymbol{\psi}(t_f)\right]^{\mathrm{T}}d\mathbf{x}(t_f)-\left\{\boldsymbol{\psi}^{\mathrm{T}}(t_f)\,\mathbf{f}\left[\mathbf{x}(t_f),\mathbf{u}(t_f)\right]-f_0\left[\mathbf{x}(t_f),\mathbf{u}(t_f)\right]\right\}dt_f$$
$$-\int_0^{t_f}\left[H(\boldsymbol{\psi},\mathbf{x},\,\mathbf{u}+\delta\mathbf{u})-H(\boldsymbol{\psi},\mathbf{x},\mathbf{u})\right]dt\,.$$

The first term is cancelled by imposing to $\psi_i(t)$ the same conditions as in the problem with fixed final time, equations (3.14) or (3.21) in the case of partially constrained final state.

The cancellation of the integral term leads to the usual condition of the maximum of H.

The only term remaining to cancel is thus the second one. Now:

$$-f_0\left[\mathbf{x}(t_f),\mathbf{u}(t_f)\right]+\boldsymbol{\psi}^{\mathrm{T}}(t_f)\cdot\mathbf{f}\left[\mathbf{x}(t_f),\mathbf{u}(t_f)\right]=H\left[\boldsymbol{\psi}(t_f),\mathbf{x}(t_f),\mathbf{u}(t_f)\right].$$

The optimal t_f value will thus correspond to

$$H\left[\psi(t_f),\mathbf{x}(t_f),\mathbf{u}(t_f)\right]=0\ . \tag{3.24}$$

This equation represents the supplementary condition required to determine the supplementary unknown t_f.

3.5 Linear Quadratic Control (LQC)

This control law, also called Linear Quadratic Controller, or LQ Controller, is well known under its abbreviation LQC. The apparent contradiction between these two adjectives will be clarified hereafter.

3.5.1 Problem Formulation

The optimal control theory presented previously is now applied to *linear*, not necessarily time-invariant, systems:

$$\begin{cases}\dot{\mathbf{x}}(t)=\mathbf{A}(t)\,\mathbf{x}(t)+\mathbf{B}(t)\,\mathbf{u}(t)\\ \mathbf{y}(t)=\mathbf{C}(t)\,\mathbf{x}(t)\end{cases}, \tag{3.25}$$

with a cost functional, or criterion, of *quadratic* nature, as seen in Sect. 3.2.1:

$$J=\frac{1}{2}\mathbf{x}^{\mathrm{T}}(t_f)\,\mathbf{S}(t)\,\mathbf{x}(t_f)+\frac{1}{2}\int_{t_0}^{t_f}\left[\mathbf{x}^{\mathrm{T}}(t)\,\mathbf{Q}(t)\,\mathbf{x}(t)+\mathbf{u}^{\mathrm{T}}(t)\,\mathbf{R}(t)\,\mathbf{u}(t)\right]dt\,, \tag{3.26}$$

The time dependencies of the elements involved in these two formulae will be omitted temporarily to simplify the writings, and will be restored after the intermediate mathematical developments.

A few comments can be made about the criterion *J*. Its second term will serve to minimize the energy and the trajectory, as said in Sect. 3.2.1.

R is a symmetric, *positive definite* matrix, since it represents a positive definite quadratic form, which means (see Sect. B.6, Appendix B) that:

$$\mathbf{u}^{\mathrm{T}}\mathbf{R}\,\mathbf{u}>0,\ \forall\,\mathbf{u}\neq\mathbf{0}\,;\quad \mathbf{u}^{\mathrm{T}}\mathbf{R}\,\mathbf{u}=0\ \text{ for }\ \mathbf{u}=0\text{ only}\,.$$

This appears clearly in the relatively frequent situation of a diagonal matrix **R**, by considering that all the terms in u_i^2 of the quadratic form representing the influence of the energy consumption are of course weighted positively.

The $\mathbf{Q}$ matrix, on the other hand, is a symmetric, *positive semidefinite* matrix. It represents thus a positive semidefinite quadratic form:

$$\mathbf{x}^{\mathrm{T}}\mathbf{Q}\,\mathbf{x} \geq 0,\ \ \forall\,\mathbf{x}\,; \quad \Rightarrow \quad \mathbf{x}^{\mathrm{T}}\mathbf{Q}\,\mathbf{x} \text{ can be zero for } \mathbf{x} \neq 0\,.$$

If indeed we consider the system output, $\mathbf{y} = \mathbf{C}\mathbf{x}$, and a weighting of the trajectory followed by this quantity by means of a positive definite quadratic form of matrix $\mathbf{G}$,

$$\mathbf{y}^{\mathrm{T}}\mathbf{G}\,\mathbf{y} = \mathbf{x}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}\mathbf{G}\mathbf{C}\,\mathbf{x} = \mathbf{x}^{\mathrm{T}}\mathbf{Q}\,\mathbf{x}\,,$$

where $\mathbf{Q} = \mathbf{C}^{\mathrm{T}}\mathbf{G}\mathbf{C}$, this quadratic form can vanish without $\mathbf{x}$ being zero, according to the shape of the $\mathbf{C}$ matrix. As an example, assume that the two first columns of $\mathbf{C}$ are identical; the nullity of $\mathbf{y}$ and thus of $\mathbf{x}^{\mathrm{T}}\mathbf{Q}\,\mathbf{x}$ will be ensured for arbitrary $x_2 = -x_1$ and $x_3 = \cdots = x_n = 0$.

The role of the first term of (3.26) is explained as follows. The usually wished final state is $\mathbf{x}(t_f) = \mathbf{x}_f = 0$, the origin of state space, or it is gained by a change of variable. It can be shown that this would lead to a physically unrealizable state feedback controller, since its parameters should grow towards infinity for $t \rightarrow t_f$.

Therefore we will have to content ourselves with bringing the system, in final state, in the vicinity of the origin, i.e. in making $\left|\mathbf{x}(t_f)\right|$ sufficiently small. This approach corresponds by the way to the physical reality, in the sense that the system parameters are never known *exactly*, and that we will never be absolutely sure that the system is at the state space origin in its final state, but only in its vicinity. This will be obtained by letting the final state free, but choosing a symmetric, *positive semidefinite* matrix $\mathbf{S}$ in an appropriate way, e.g. in the form of a diagonal matrix whose elements are much larger than those of the matrices $\mathbf{Q}$ and $\mathbf{R}$. The quantities weighted by high coefficients will then necessarily become small when J reaches its minimum.

The final state being now free, the transversality condition which applies here is equation (3.14):

$$\boldsymbol{\psi}(t_f) = -\left.\frac{dh}{d\mathbf{x}}\right|_{t=t_f}.$$

Since, in the present case, $h\left[\mathbf{x}(t_f)\right] = (1/2)\,\mathbf{x}^{\mathrm{T}}\mathbf{S}\,\mathbf{x}$, this yields, with (B.10):

$$\boldsymbol{\psi}(t_f) = -\mathbf{S}\,\mathbf{x}(t_f)\,.$$

Summary. The present problem is a particular case of the basic problem with free final state, with a cost functional given by the quadratic criterion (3.26), accompanied by the following boundary conditions:

$$\mathbf{x}(0)=\mathbf{x}_0\,,$$

$$\text{given } t_f>0,\ \ \boldsymbol{\psi}(t_f)=-\mathbf{S}\,\mathbf{x}(t_f)\,. \tag{3.27}$$

Remark 3.7. Note that in a more general framework, including among others the case of time-varying systems, the **Q**, **R** and **S** matrices can very well depend on time, exactly as the matrices **A**, **B** and **C** describing the plant.

3.5.2 Optimal Control Law. Riccati Equation

From the comparison of (3.26) with (3.2), we deduce that here

$$f_0(\mathbf{x},\mathbf{u},t)=\frac{1}{2}(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\,\mathbf{x}+\mathbf{u}^{\mathrm{T}}\mathbf{R}\,\mathbf{u})\,. \tag{3.28}$$

Therefore, according to (3.4),

$$\begin{aligned}H(\boldsymbol{\psi},\mathbf{x},\mathbf{u})&=-\frac{1}{2}(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}+\mathbf{u}^{\mathrm{T}}\mathbf{R}\,\mathbf{u})+\boldsymbol{\psi}^{\mathrm{T}}(\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{u})\\&=-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}-\frac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{R}\,\mathbf{u}+\boldsymbol{\psi}^{\mathrm{T}}\mathbf{A}\mathbf{x}+\boldsymbol{\psi}^{\mathrm{T}}\mathbf{B}\mathbf{u}\,.\end{aligned}$$

Thus, with the differentiation rules (B.10) and (B.8), the following holds:

$$\frac{\partial H}{\partial\,\mathbf{x}}=-\,\mathbf{Q}\mathbf{x}+\mathbf{A}^{\mathrm{T}}\boldsymbol{\psi}\,.$$

According to (3.13), the resulting adjoint equation is

$$\dot{\boldsymbol{\psi}}=-\mathbf{A}^{\mathrm{T}}\boldsymbol{\psi}+\mathbf{Q}\,\mathbf{x}\,. \tag{3.29}$$

The value of **u** which makes *H* maximum is given by

$$\frac{\partial H}{\partial\mathbf{u}}=0\,,$$

thus, according to the same differentiation rules as above,

$$\frac{\partial H}{\partial \mathbf{u}} = -\mathbf{R}\,\mathbf{u} + \mathbf{B}^{\mathrm{T}}\boldsymbol{\psi} = 0\,.$$

The sought optimal control law, which will be denoted by $\mathbf{u}^{*}(t)$, is thus:

$$\mathbf{u}(t) = \mathbf{u}^{*}(t) = \mathbf{R}^{-1}(t)\,\mathbf{B}^{\mathrm{T}}(t)\,\boldsymbol{\psi}(t)\ . \tag{3.30}$$

Remark 3.8. The matrix $\mathbf{R}^{-1}$ exists, since **R** has been assumed positive definite, thus regular.

By substitution of (3.30) into the state equation (3.25), we have

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\,\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\boldsymbol{\psi}\,. \tag{3.31}$$

Equations (3.29) and (3.31) constitute the *Hamilton's equations* of the system. Since they are linear in **x** and in $\boldsymbol{\psi}$, it is foreseeable that $\boldsymbol{\psi}(t)$ will depend linearly on $\mathbf{x}(t)$. It is thus logic to look for a solution of the form

$$\boldsymbol{\psi}(t) = -\,\mathbf{P}(t)\,\mathbf{x}(t)\,, \tag{3.32}$$

where $\mathbf{P}(t)$ is an unknown $(n \times n)$ matrix. Inserting (3.32) into the system of equations (3.31) and (3.29), we obtain

$$\dot{\mathbf{x}} = (\mathbf{A} - \mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P})\,\mathbf{x}\,, \tag{3.33}$$

and

$$-\dot{\mathbf{P}}\mathbf{x} - \mathbf{P}\dot{\mathbf{x}} = (\mathbf{A}^{\mathrm{T}}\mathbf{P} + \mathbf{Q})\,\mathbf{x}\ . \tag{3.34}$$

The left multiplication of the first of these two equations by **P** followed by its addition to the second one yields

$$-\,\dot{\mathbf{P}}\mathbf{x} = (\mathbf{P}\mathbf{A} + \mathbf{A}^{\mathrm{T}}\mathbf{P} - \mathbf{P}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P} + \mathbf{Q})\,\mathbf{x}\,.$$

Since this relation should hold whatever **x**, the following holds:

$$-\,\dot{\mathbf{P}}(t) = \mathbf{P}(t)\mathbf{A}(t) + \mathbf{A}^{\mathrm{T}}(t)\mathbf{P}(t) - \mathbf{P}(t)\mathbf{B}(t)\mathbf{R}^{-1}(t)\mathbf{B}^{\mathrm{T}}(t)\mathbf{P}(t) + \mathbf{Q}(t)\ , \tag{3.35}$$

where the full time dependencies have been restored.

The final condition $\boldsymbol{\psi}(t_f) = -\mathbf{S}\,\mathbf{x}(t_f)$ applied to (3.32) has the consequence that

$$\mathbf{P}(t_f) = \mathbf{S} \ . \tag{3.36}$$

The first order differential equation (3.35), which is quadratic in the unknown $\mathbf{P}(t)$, is called *Riccati (matrix) differential equation*. Since it is quadratic, it has several solutions $\mathbf{P}(t)$.

Any solution $\mathbf{P}(t)$ is symmetric: indeed, it is easy to see that, since the matrices $\mathbf{Q}(t)$ and $\mathbf{R}(t)$ are symmetric for any t, $\mathbf{P}^{\mathrm{T}}(t)$ is a solution of (3.35) if and only if $\mathbf{P}(t)$ is symmetric. We will prove in Sect. 3.5.4 that one of the solutions is positive semidefinite, or positive definite, according to the situation (see corresponding criteria in Appendix B).

The solution of (3.35) is generally obtained by numerical integration, integrating backwards in time from $t = t_f$ with the *initial* condition, denoted here rightly a *final* condition, given by (3.36).

3.5.3 State-feedback Optimal Control

The state feedback solution to the optimal control problem of a linear system by quadratic criterion is obtained by combining (3.30) and (3.32):

$$\mathbf{u}^*(t) = -\mathbf{R}^{-1}(t)\,\mathbf{B}^{\mathrm{T}}(t)\,\mathbf{P}(t)\,\mathbf{x}(t)\,, \tag{3.37}$$

thus:

$$\mathbf{u}(t) = \mathbf{u}^*(t) = -\mathbf{L}(t)\mathbf{x}(t) \ , \tag{3.38}$$

$$\mathbf{L}(t) = \mathbf{R}^{-1}(t)\,\mathbf{B}^{\mathrm{T}}(t)\,\mathbf{P}(t) \ , \tag{3.39}$$

where $\mathbf{P}(t)$ is given by (3.35).

The $(p \times n)$ matrix $\mathbf{L}(t)$ can be calculated before starting the system. Then along its evolution, the control $\mathbf{u}(t)$ is calculated at each time from the present state by (3.38), which is a state feedback control.

3.5.4 Cost Functional Minimum and Positive Semidefiniteness of P

Let us introduce the optimal control law (3.37) in the quadratic criterion expression (3.26). Calculate first the second quadratic form:

$$\mathbf{u}^{*\mathrm{T}}\mathbf{R}\,\mathbf{u}^{*}=(-\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}\,\mathbf{x})^{\mathrm{T}}\,\mathbf{R}\,(-\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}\,\mathbf{x})=\mathbf{x}^{\mathrm{T}}\mathbf{P}\mathbf{B}(\mathbf{R}^{-1})^{\mathrm{T}}\,\mathbf{R}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}\,\mathbf{x}\,,$$

since **P** is symmetric. Since **R** is also symmetric, $(\mathbf{R}^{-1})^{\mathrm{T}}=(\mathbf{R}^{\mathrm{T}})^{-1}=\mathbf{R}^{-1}$, and

$$\mathbf{u}^{*\mathrm{T}}\mathbf{R}\,\mathbf{u}^{*}=\mathbf{x}^{\mathrm{T}}\mathbf{P}\,\mathbf{B}\,\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}\,\mathbf{x}\,. \tag{3.40}$$

Let us show then that, in this case, the integrand involved in the corresponding cost functional (3.26), thus in its minimum value $J_{\min}$, has the following expression:

$$\mathbf{x}^{\mathrm{T}}\mathbf{Q}\,\mathbf{x}+\mathbf{u}^{*\mathrm{T}}\mathbf{R}\,\mathbf{u}^{*}=-\frac{d}{dt}\Big[\mathbf{x}^{\mathrm{T}}(t)\mathbf{P}(t)\mathbf{x}(t)\Big]\,. \tag{3.41}$$

To do this, write

$$\frac{d}{dt}(\mathbf{x}^{\mathrm{T}}\mathbf{P}\mathbf{x})=\dot{\mathbf{x}}^{\mathrm{T}}\mathbf{P}\,\mathbf{x}+\mathbf{x}^{\mathrm{T}}\dot{\mathbf{P}}\,\mathbf{x}+\mathbf{x}^{\mathrm{T}}\mathbf{P}\,\dot{\mathbf{x}}=\dot{\mathbf{x}}^{\mathrm{T}}\mathbf{P}\,\mathbf{x}+\mathbf{x}^{\mathrm{T}}(\dot{\mathbf{P}}\,\mathbf{x}+\mathbf{P}\,\dot{\mathbf{x}})\,.$$

By using the transpose of (3.33) as well as (3.34), we obtain successively:

$$\begin{aligned}\frac{d}{dt}(\mathbf{x}^{\mathrm{T}}\mathbf{P}\mathbf{x})&=\mathbf{x}^{\mathrm{T}}(\mathbf{A}^{\mathrm{T}}-\mathbf{P}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}})\mathbf{P}\,\mathbf{x}-\mathbf{x}^{\mathrm{T}}(\mathbf{A}^{\mathrm{T}}\mathbf{P}+\mathbf{Q})\mathbf{x}\\&=\mathbf{x}^{\mathrm{T}}(\mathbf{A}^{\mathrm{T}}\mathbf{P}-\mathbf{P}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}-\mathbf{A}^{\mathrm{T}}\mathbf{P}-\mathbf{Q})\,\mathbf{x}\\&=-\,\mathbf{x}^{\mathrm{T}}(\mathbf{P}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P})\,\mathbf{x}-\mathbf{x}^{\mathrm{T}}\mathbf{Q}\,\mathbf{x}=-\mathbf{u}^{*\mathrm{T}}\mathbf{R}\,\mathbf{u}^{*}-\mathbf{x}^{\mathrm{T}}\mathbf{Q}\,\mathbf{x}\,,\end{aligned}$$

after taking into account (3.40). By integrating now (3.41) from *t* to t_f and accounting for the final condition (3.36), we get:

$$\mathbf{x}^{\mathrm{T}}(t)\mathbf{P}(t)\mathbf{x}(t)=\mathbf{x}^{\mathrm{T}}(t_f)\,\mathbf{S}\,\mathbf{x}(t_f)+\int_{t}^{t_f}(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\,\mathbf{x}+\mathbf{u}^{*\mathrm{T}}\mathbf{R}\,\mathbf{u}^{*})\,d\tau\,. \tag{3.42}$$

By letting $t=t_0$ in this expression and comparing the result with (3.26), the following minimum cost functional is obtained:

$$J_{\min}=\frac{1}{2}\,\mathbf{x}^{\mathrm{T}}(t_0)\,\mathbf{P}(t_0)\,\mathbf{x}(t_0)\ \ . \tag{3.43}$$

A detailed examination of (3.42) allows to draw conclusions on the *sign* of the matrix **P**. The quadratic form in **S** is positive semidefinite by hypothesis. As to the integral on the right side of the relation, two situations should be considered.

First case. The plant to control is not completely observable.

Its observability canonical decomposition, assuming that $\mathbf{x}_2$ represents the unobservable part of the state, is then the following (Appendix A, Sect. A.3.3):

$$\begin{cases} \begin{pmatrix} \dot{\mathbf{x}}_1 \\ \dot{\mathbf{x}}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{11} & 0 \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} + \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix} \mathbf{u} \\ \quad \mathbf{y} = \begin{pmatrix} \mathbf{C}_1 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \end{cases}$$

The return to equilibrium, from an initial state $\mathbf{x}^{\mathrm{T}}(t) = (\mathbf{0}^{\mathrm{T}} \quad \mathbf{x}_{2,t}^{\mathrm{T}})$ at initial time t, of this system equipped with a control law of the form $\mathbf{u}(\tau) = -\mathbf{L}\mathbf{x}(\tau)$, with $\mathbf{L} = (\mathbf{L}_1 \quad \mathbf{0})$, thus with the system matrix

$$\mathbf{A}_{CL} = \mathbf{A} - \mathbf{B}\mathbf{L} = \begin{pmatrix} \mathbf{A}_{11} - \mathbf{B}_1\mathbf{L}_1 & 0 \\ \mathbf{A}_{21} - \mathbf{B}_2\mathbf{L}_1 & \mathbf{A}_{22} - \mathbf{B}_2\mathbf{L}_2 \end{pmatrix},$$

is given by:

$$\begin{cases} \mathbf{x}_1(\tau) \equiv 0, \ \forall \tau \in \left[t, t_f\right], \text{ since } \dot{\mathbf{x}}_1(\tau) = (\mathbf{A}_{11} - \mathbf{B}_1\mathbf{L}_1)\,\mathbf{x}_1(\tau) = 0, \forall \tau \\ \mathbf{x}_2(\tau) = \exp\left[(\mathbf{A}_{22} - \mathbf{B}_2\mathbf{L}_2)(\tau - t)\right] \mathbf{x}_{2,t} \end{cases}$$

The control law amounts then to $\mathbf{u}(\tau) \equiv 0, \forall\, \tau \in \left[t, t_f\right]$. If we choose now $\mathbf{Q} = \mathbf{C}^{\mathrm{T}}\mathbf{C}$, which yields $\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} = \mathbf{x}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}\mathbf{C}\mathbf{x} = \mathbf{y}^{\mathrm{T}}\mathbf{y} = 0, \forall \tau$, since $\mathbf{y}(\tau) = \mathbf{C}_1\mathbf{x}_1(\tau) \equiv 0, \forall \tau$, the integral on the right side of (3.42) vanishes. This means that the right side can vanish for $\mathbf{x}(t) \neq 0$. The quadratic form on the left side is thus *positive semidefinite*, therefore so is matrix $\mathbf{P}(t)$.

Second case. The plant to control is completely observable.

The reasoning differs from the one of the previous case as to the following points: the control law for the return to equilibrium $\mathbf{u}(\tau)$ is no longer identically zero, and the integral of the right side will be strictly positive for any $t < t_f$, since it was assumed that $\mathbf{Q}$ is positive semidefinite and that $\mathbf{R}$ is positive definite. As a result, $\mathbf{x}^{\mathrm{T}}(t)\mathbf{P}(t)\mathbf{x}(t) > 0$ for $0 \le t < t_f$, which means that $\mathbf{P}(t)$ is in this case *positive definite* for $0 \le t < t_f$.

3.5.5 Solving the Riccati Equation. Hamiltonian Matrix.

3.5.5.1 Hamiltonian System and Hamiltonian Matrix

Let us start from the Hamilton's equations (3.31) and (3.29) of the system:

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\boldsymbol{\psi} \\ \dot{\boldsymbol{\psi}} = \mathbf{Q}\mathbf{x} - \mathbf{A}^{\mathrm{T}}\boldsymbol{\psi} \end{cases}$$

We may rewrite these equations as the homogeneous equation of an augmented system, which has the state vector $(\mathbf{x}^{\mathrm{T}}, \boldsymbol{\psi}^{\mathrm{T}})^{\mathrm{T}}$:

$$\begin{pmatrix} \dot{\mathbf{x}} \\ \dot{\boldsymbol{\psi}} \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}} \\ \mathbf{Q} & -\mathbf{A}^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \boldsymbol{\psi} \end{pmatrix} = \mathbf{H} \begin{pmatrix} \mathbf{x} \\ \boldsymbol{\psi} \end{pmatrix}. \tag{3.44}$$

This system is called *Hamiltonian linear system*, of order $2n$, and the matrix

$$\mathbf{H} = \begin{pmatrix} \mathbf{A} & \mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}} \\ \mathbf{Q} & -\mathbf{A}^{\mathrm{T}} \end{pmatrix} \tag{3.45}$$

is the corresponding $(2n \times 2n)$ *Hamiltonian matrix.*

3.5.5.2 Solution of the Riccati Equation

Let $\boldsymbol{\Phi}(t, t_f)$ be the $(2n \times 2n)$ transition matrix of this system. We know that it is itself a solution of the homogeneous differential equation, and can thus in theory be calculated by solving

$$\dot{\boldsymbol{\Phi}}(t, t_f) = \mathbf{H}\,\boldsymbol{\Phi}(t, t_f), \tag{3.46}$$

with the final condition

$$\boldsymbol{\Phi}(t_f, t_f) = \mathbf{I}_{2n}. \tag{3.47}$$

The solution of (3.44) is then given, starting from the final state $\mathbf{x}(t_f)$, by

$$\begin{pmatrix} \mathbf{x}(t) \\ \boldsymbol{\psi}(t) \end{pmatrix} = \mathbf{\Phi}(t,t_f) \begin{pmatrix} \mathbf{x}(t_f) \\ \boldsymbol{\psi}(t_f) \end{pmatrix}. \tag{3.48}$$

This equation can be rewritten in partitioned form as

$$\begin{pmatrix} \mathbf{x}(t) \\ \boldsymbol{\psi}(t) \end{pmatrix} = \begin{pmatrix} \mathbf{\Phi}_{11}(t,t_f) & \mathbf{\Phi}_{12}(t,t_f) \\ \mathbf{\Phi}_{21}(t,t_f) & \mathbf{\Phi}_{22}(t,t_f) \end{pmatrix} \begin{pmatrix} \mathbf{x}(t_f) \\ \boldsymbol{\psi}(t_f) \end{pmatrix}, \tag{3.49}$$

where the four blocks $\mathbf{\Phi}_{ij}(t,t_f)$ are of size $(n \times n)$. This yields

$$\begin{aligned} \mathbf{x}(t) &= \mathbf{\Phi}_{11}(t,t_f)\mathbf{x}(t_f) + \mathbf{\Phi}_{12}(t,t_f)\boldsymbol{\psi}(t_f) \\ \boldsymbol{\psi}(t) &= \mathbf{\Phi}_{21}(t,t_f)\mathbf{x}(t_f) + \mathbf{\Phi}_{22}(t,t_f)\boldsymbol{\psi}(t_f) \end{aligned}$$

With the use of the transversality condition (3.27), namely $\boldsymbol{\psi}(t_f) = -\mathbf{S}\,\mathbf{x}(t_f)$, the two previous equations become

$$\begin{aligned} \mathbf{x}(t) &= \left[\mathbf{\Phi}_{11}(t,t_f) - \mathbf{\Phi}_{12}(t,t_f)\,\mathbf{S}\right]\mathbf{x}(t_f) \\ \boldsymbol{\psi}(t) &= \left[\mathbf{\Phi}_{21}(t,t_f) - \mathbf{\Phi}_{22}(t,t_f)\,\mathbf{S}\right]\mathbf{x}(t_f) \end{aligned}$$

from where it results, by elimination of $\mathbf{x}(t_f)$, that

$$\boldsymbol{\psi}(t) = \left[\mathbf{\Phi}_{21}(t,t_f) - \mathbf{\Phi}_{22}(t,t_f)\,\mathbf{S}\right]\left[\mathbf{\Phi}_{11}(t,t_f) - \mathbf{\Phi}_{12}(t,t_f)\,\mathbf{S}\right]^{-1}\mathbf{x}(t).$$

By comparison with (3.32), the solution of the Riccati equation is obtained in the form:

$$\mathbf{P}(t) = -\left[\mathbf{\Phi}_{21}(t,t_f) - \mathbf{\Phi}_{22}(t,t_f)\,\mathbf{S}\right]\left[\mathbf{\Phi}_{11}(t,t_f) - \mathbf{\Phi}_{12}(t,t_f)\,\mathbf{S}\right]^{-1}. \tag{3.50}$$

3.5.5.3 Some Properties of the Hamiltonian Matrix

- **Pure imaginary matrix:** $\mathbf{J} = \begin{pmatrix} \mathbf{0} & \mathbf{I}_n \\ -\mathbf{I}_n & \mathbf{0} \end{pmatrix}$.

It is the extension of $j = \sqrt{-1}$ to $2n$ dimensions. Indeed, $\mathbf{J}^2 = -\mathbf{I}_{2n}$, as readily verified. One will verify also that $\mathbf{J}^{\mathrm{T}} = -\mathbf{J} = \mathbf{J}^{-1}$.

- **Property 1 of the Hamiltonian matrix**: $\mathbf{J\,HJ} = \mathbf{H}^{\mathrm{T}}$.

Proof. By taking into account the symmetry of **Q** and of **R**,

$$\mathbf{JHJ} = \begin{pmatrix} \mathbf{0} & \mathbf{I}_n \\ -\mathbf{I}_n & \mathbf{0} \end{pmatrix}\begin{pmatrix} \mathbf{A} & \mathbf{BR}^{-1}\mathbf{B}^{\mathrm{T}} \\ \mathbf{Q} & -\mathbf{A}^{\mathrm{T}} \end{pmatrix}\begin{pmatrix} \mathbf{0} & \mathbf{I}_n \\ -\mathbf{I}_n & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{Q} & -\mathbf{A}^{\mathrm{T}} \\ -\mathbf{A} & -\mathbf{BR}^{-1}\mathbf{B}^{\mathrm{T}} \end{pmatrix}\begin{pmatrix} \mathbf{0} & \mathbf{I}_n \\ -\mathbf{I}_n & \mathbf{0} \end{pmatrix}$$
$$= \begin{pmatrix} \mathbf{A}^{\mathrm{T}} & \mathbf{Q} \\ \mathbf{BR}^{-1}\mathbf{B}^{\mathrm{T}} & -\mathbf{A} \end{pmatrix} = \mathbf{H}^{\mathrm{T}} \quad .$$

- **Property 2 of the Hamiltonian matrix**: if λ is an eigenvalue of **H**, then $-\lambda$ is also eigenvalue of **H**.

Proof. From the previous properties it follows that $\mathbf{J}^{-1}\mathbf{HJ} = -\mathbf{JHJ} = -\mathbf{H}^{\mathrm{T}}$. The matrices **H** and $-\mathbf{H}^{\mathrm{T}}$ correspond thus to each other in a similitude transformation, and, on account of that, have the same eigenvalues (Appendix A, Sect. A.7.1). Therefore, if λ is an eigenvalue of **H**, $-\lambda$, which is trivially eigenvalue of $-\mathbf{H}$ thus of $-\mathbf{H}^{\mathrm{T}}$, is also eigenvalue of **H.**

3.5.6 Case of the Linear Time-invariant (LTI) Systems

Assume now constant all the matrices which are involved in the problem definition (time-invariant plant, cost functional quadratic forms without explicit time dependence): $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{Q}, \mathbf{R}, \mathbf{S} = \text{ constant}$.

The optimal controller

$$\mathbf{u} = -\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}(t)\,\mathbf{x} = -\mathbf{L}(t)\,\mathbf{x}$$

is *not* time-invariant, since the Riccati equation solution $\mathbf{P}(t)$ still depends on time.

Indeed, even though the Hamiltonian matrix is in that case constant, as shows its definition (3.45), the transition matrix $\mathbf{\Phi}(t,t_f)$ of the corresponding Hamiltonian system, which depends here only on the time interval separating the two instants,

$$\mathbf{\Phi}(t,t_f) = e^{\mathbf{H}(t-t_f)} = e^{-\mathbf{H}\tau}, \qquad (3.51)$$

where $\tau = t_f - t$, remains a function of time, thus also the $\mathbf{\Phi}_{ij}(t,t_f)$, and consequently also **P** according to (3.50).

Infinite horizon stationary solution. In order to obtain a time independent controller, thus easier to implement, one introduces the following additional hypothesis:

$$t_f \to \infty \, .$$

For the integral in the cost functional J to converge, it is then necessary that:

$$\lim_{t \to \infty} \mathbf{x}(t) = 0 \, ,$$

from where follows the final condition:

$$\mathbf{x}(t_f) = \mathbf{x}(\infty) = 0 \, . \tag{3.52}$$

This cancels the quadratic form in $\mathbf{x}(t_f)$, with associated matrix **S**. It disappears therefore from J, which reduces to:

$$J = \frac{1}{2} \int_0^\infty \left[\mathbf{x}^\mathrm{T}(t)\, \mathbf{Q}\, \mathbf{x}(t) + \mathbf{u}^\mathrm{T}(t)\, \mathbf{R}\, \mathbf{u}(t) \right] dt \, . \tag{3.53}$$

The controller so designed for $t_f \to \infty$, and used instead of the optimal controller, will give to the system in *finite* time only an *approximately optimal behavior*, called *suboptimal*, but sufficiently close to optimal for most practical applications.

With the hypothesis $t_f \to \infty$, the backward integrated solution to the Riccati differential equation must tend, for $t \to 0$, towards a constant matrix. [AnMo89], [ZhDG96] and [Duc01], [Duc04] have shown that, for that limit to exist and to lead to a stable closed-loop system, it is necessary that the following conditions be satisfied:

1. the (**A**,**B**) pair is stabilizable;
2. the $(\mathbf{Q}_0, \mathbf{A})$ pair is detectable, where $\mathbf{Q}_0$ is any matrix such that $\mathbf{Q} = \mathbf{Q}_0^\mathrm{T} \mathbf{Q}_0$.

Consequently, at the limit, $\dot{\mathbf{P}} \to 0$ and the differential Riccati equation (3.35) is replaced by the *algebraic Riccati equation* (ARE) for *continuous-time systems*:

$$\mathbf{PA} + \mathbf{A}^\mathrm{T}\mathbf{P} - \mathbf{PBR}^{-1}\mathbf{B}^\mathrm{T}\mathbf{P} + \mathbf{Q} = \mathbf{0} \, , \tag{3.54}$$

which is a nonlinear algebraic matrix equation, having thus several solutions **P**, which are constant.

The unique *positive semidefinite* solution $\mathbf{P} = \mathbf{P}^{\mathrm{T}}$ of this equation is also the only stabilizing solution, i.e. the one which ensures the stability of the system fed back by the optimal state feedback matrix (3.39), which becomes then also a constant matrix amounting to:

$$\mathbf{L} = \mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P} \ . \tag{3.55}$$

In the cited works it is also shown that, if in addition to the previous hypotheses the $(\mathbf{Q}_0, \mathbf{A})$ pair is completely observable, the algebraic Riccati equation has a unique *positive definite* solution, which is also *the* stabilizing solution.

To solve (3.54), there are two methods.

Direct solving. As a matter of fact, one solves the numerical equivalent of (3.35), which will be presented in Sect. 3.6 dealing with discrete-time systems, by looking for the stationary solution towards which $\mathbf{P}(k)$ converges.

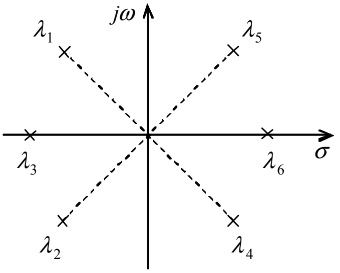
Solving by the Hamiltonian matrix. In the Hamilton's equations (3.29) and (3.31) of the system, $\mathbf{x}$ is the state vector of the *closed-loop* system containing the optimal controller, and $\boldsymbol{\psi}$ the costate vector.

The corresponding characteristic equation

$$\det(\lambda \mathbf{I}_{2n} - \mathbf{H}) = 0$$

yields the eigenvalues of the Hamiltonian system. It was established in Sect. 3.5.5.3 that, if λ is an eigenvalue, then $-\lambda$ is one also. If the conditions 1 and 2 above are satisfied, there are no eigenvalues on the imaginary axis, as shown in [ZhDG96].

Let us number from 1 to n the eigenvalues which lie to the left of the imaginary axis in the complex s-plane: $\lambda_1, \ldots, \lambda_n$, and from $n+1$ to $2n$ those which lie to the right of this axis: $\lambda_{n+1}, \ldots, \lambda_{2n}$, as represented schematically on the right for an imaginary situation with $n = 3$.



The free response of the Hamiltonian system can be decomposed in its *eigenmodes* as follows (see equation (A.45) in Appendix A):

$$\begin{pmatrix} \mathbf{x}(t) \\ \boldsymbol{\psi}(t) \end{pmatrix} = \sum_{i=1}^{2n} c_i \, e^{\lambda_i t} \begin{pmatrix} \mathbf{v}_{ix} \\ \mathbf{v}_{i\psi} \end{pmatrix},$$

where $\begin{pmatrix} \mathbf{v}_{ix}^{\mathrm{T}} & \mathbf{v}_{i\psi}^{\mathrm{T}} \end{pmatrix}^{\mathrm{T}}$ is the eigenvector associated with the eigenvalue λ_i.

Remarking that the eigenmodes associated with the eigenvalues $\lambda_{n+1},\ldots,\lambda_{2n}$ are divergent, whereas it is expected that $\mathbf{x}(\infty)=0$, it is seen that, necessarily, $c_{n+1}=\ldots=c_{2n}=0$. Hence:

$$\begin{pmatrix}\mathbf{x}(t)\\ \boldsymbol{\psi}(t)\end{pmatrix}=\begin{pmatrix}\mathbf{v}_{1x}\\ \mathbf{v}_{1\psi}\end{pmatrix}c_1\,e^{\lambda_1 t}+\ldots+\begin{pmatrix}\mathbf{v}_{nx}\\ \mathbf{v}_{n\psi}\end{pmatrix}c_n\,e^{\lambda_n t}=\begin{pmatrix}\mathbf{v}_{1x} & \cdots & \mathbf{v}_{nx}\\ \mathbf{v}_{1\psi} & \cdots & \mathbf{v}_{n\psi}\end{pmatrix}\begin{pmatrix}c_1\,e^{\lambda_1 t}\\ \vdots\\ c_n\,e^{\lambda_n t}\end{pmatrix},$$

which yields

$$\mathbf{x}(t)=\begin{pmatrix}\mathbf{v}_{1x} & \cdots & \mathbf{v}_{nx}\end{pmatrix}\begin{pmatrix}c_1\,e^{\lambda_1 t} & \cdots & c_n\,e^{\lambda_n t}\end{pmatrix}^{\mathrm{T}}, \tag{3.56}$$

$$\boldsymbol{\psi}(t)=\begin{pmatrix}\mathbf{v}_{1\psi} & \cdots & \mathbf{v}_{n\psi}\end{pmatrix}\begin{pmatrix}\mathbf{v}_{1x} & \cdots & \mathbf{v}_{nx}\end{pmatrix}^{-1}\mathbf{x}(t),$$

and, finally,

$$\mathbf{u}(t)=\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\begin{pmatrix}\mathbf{v}_{1\psi} & \cdots & \mathbf{v}_{n\psi}\end{pmatrix}\begin{pmatrix}\mathbf{v}_{1x} & \cdots & \mathbf{v}_{nx}\end{pmatrix}^{-1}\mathbf{x}(t)\ . \tag{3.57}$$

Remark 3.9. The eigenvalues $\lambda_{L1},\ldots,\lambda_{Ln}$ of the closed loop equipped with this optimal control are exactly the *n* eigenvalues of the Hamiltonian system lying at the left of the imaginary axis. This stems directly from (3.56), $\mathbf{x}(t)$ being the closed-loop state vector, since according to (3.44) and (3.32),

$$\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\boldsymbol{\psi}=(\mathbf{A}-\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P})\mathbf{x}\,.$$

Remark 3.10. Link with the complete modal control (Sect. 1.6).
[Rop90] has shown that the above optimal control can be seen as a complete modal design, in which the eigenvalues $\lambda_{L1},\ldots,\lambda_{Ln}$ prescribed to the closed-loop are to be taken equal to the *n* solutions of the following equation:

$$\det\mathbf{M}(s)=0\,, \tag{3.58}$$

where

$$\mathbf{M}(s)=\mathbf{I}+\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}(-s\mathbf{I}-\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{Q}\,(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\,, \tag{3.59}$$

and the invariant parameter vectors are to be taken equal to the non trivial solutions of the following *n* homogeneous equation systems:

$$\mathbf{M}(\lambda_{Li})\,\mathbf{p}_i=0,\quad i=1,\ldots,n\,. \tag{3.60}$$

Under these conditions, the control law given by the formula (1.63) of Chap. 1 is identical to the one which would result from the use of (3.54) and of (3.55) with a cost functional $J = \frac{1}{2}\int_0^\infty \left[\mathbf{x}^{\mathrm{T}}(t)\,\mathbf{Q}\,\mathbf{x}(t) + \mathbf{u}^{\mathrm{T}}(t)\,\mathbf{R}\,\mathbf{u}(t)\right] dt$.

One should note that, at the difference to a complete modal design of pole placement type, where the choice of the parameter vectors is arbitrary, these vectors are here entirely determined due to the optimality constraint imposed through the weighting matrices of the quadratic criterion **Q** and **R**.

3.6 Case of the Discrete Systems

3.6.1 Basic Problem

The case of the discrete-time systems is very similar to the one of the continuous-time systems. We will thus mention here only the main differences between these two situations.

3.6.1.1 Starting Hypotheses

For the optimal control problem the given data are here:

- a dynamic system, defined for $k = 0, 1, \dots, N-1$ by

$$\mathbf{x}_{k+1} = \mathbf{f}\left(\mathbf{x}_k, \mathbf{u}_k\right), \tag{3.61}$$

$$\mathbf{x}_0 \ \text{ given;}$$

- a cost functional, which involves the specified final time, $t_f = NT_s$ if the discrete plant results from a sampling process at a period T_s, or simply $t_f = N$ if the time progress is counted just as a number of discrete steps:

$$J = h(\mathbf{x}_N) + \sum_{k=0}^{N-1} f_0(\mathbf{x}_k, \mathbf{u}_k)\,; \tag{3.62}$$

- a control which remains constant during one sampling time or time slot

$$\mathbf{u} = \mathbf{u}_k = \text{constant, for } t \in \left[kT_s, (k+1)T_s\right].$$

Remark 3.11. The summation limitation in (3.62) at the time step $N-1$ expresses the fact that *J* must not involve controls out of the interval $\left[0, NT_s\right]$, taking into account that $\mathbf{u}_N$ is applied in the interval $\left[N, N+1\right]$.

Remark 3.12. The problem is identical to the one introduced in the continuous case by the equations (3.1) and (3.2).

3.6.1.2 Modified Cost Functional

$$J = h(\mathbf{x}_N) + \sum_{k=0}^{N-1} f_0(\mathbf{x}_k, \mathbf{u}_k) + \sum_{k=0}^{N-1} \boldsymbol{\psi}_{k+1}^{\mathrm{T}} \left[\mathbf{x}_{k+1} - \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k)\right], \tag{3.63}$$

where the sequence $\left\{\boldsymbol{\psi}_k\right\}$ is for the time being unknown. Note here the presence of $\boldsymbol{\psi}_{k+1}$, and not of $\boldsymbol{\psi}_k$, since this Lagrange multiplier must apply a term calculated at $k+1$ and not at *k*.

3.6.1.3 Hamiltonian

As in the continuous case, the *Hamiltonian*

$$H(\boldsymbol{\psi}, \mathbf{x}, \mathbf{u}) = -f_0(\mathbf{x}, \mathbf{u}) + \boldsymbol{\psi}^{\mathrm{T}} \cdot \mathbf{f}(\mathbf{x}, \mathbf{u}) \tag{3.64}$$

is introduced here, its more detailed expression being

$$H(\boldsymbol{\psi}_{k+1}, \mathbf{x}_k, \mathbf{u}_k) = -f_0(\mathbf{x}_k, \mathbf{u}_k) + \boldsymbol{\psi}_{k+1}^{\mathrm{T}} \cdot \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \ . \tag{3.65}$$

The modified cost functional writes thus

$$J = h(\mathbf{x}_N) + \sum_{k=0}^{N-1} \left[\boldsymbol{\psi}_{k+1}^{\mathrm{T}} \cdot \mathbf{x}_{k+1} - H(\boldsymbol{\psi}_{k+1}, \mathbf{x}_k, \mathbf{u}_k)\right].$$

This expression is completely similar to (3.5) and can be written also

$$J = h(\mathbf{x}_N) + \sum_{k=0}^{N-1} \left[\boldsymbol{\psi}_k^{\mathrm{T}} \cdot \mathbf{x}_k - H(\boldsymbol{\psi}_{k+1}, \mathbf{x}_k, \mathbf{u}_k)\right] + \boldsymbol{\psi}_N^{\mathrm{T}} \cdot \mathbf{x}_N - \boldsymbol{\psi}_0^{\mathrm{T}} \cdot \mathbf{x}_0 \, . \tag{3.66}$$

3.6.2 Variational Calculus

As in the continuous case, a small variation is applied to the supposedly optimal control law $\mathbf{u}_k$, and the resulting cost functional variation is evaluated:

$$\delta J = h(\mathbf{x}_N + \delta\mathbf{x}_N) - h(\mathbf{x}_N) + \boldsymbol{\psi}_N^{\mathrm{T}} \cdot \delta\mathbf{x}_N - \boldsymbol{\psi}_0^{\mathrm{T}} \cdot \delta\mathbf{x}_0$$
$$+ \sum_{k=0}^{N-1} \left[\boldsymbol{\psi}_k^{\mathrm{T}} \cdot \delta\mathbf{x}_k - H(\boldsymbol{\psi}_{k+1}, \mathbf{x}_k + \delta\mathbf{x}_k, \mathbf{u}_k + \delta\mathbf{u}_k) + H(\boldsymbol{\psi}_{k+1}, \mathbf{x}_k, \mathbf{u}_k) \right].$$

A Taylor series expansion, limited to first order approximation, yields,

$$\sum_{k=0}^{N-1} \left[H(\boldsymbol{\psi}_{k+1}, \mathbf{x}_k + \delta\mathbf{x}_k, \mathbf{u}_k + \delta\mathbf{u}_k) - H(\boldsymbol{\psi}_{k+1}, \mathbf{x}_k, \mathbf{u}_k) \right]$$
$$\simeq \sum_{k=0}^{N-1} \left[\frac{\partial}{\partial \mathbf{x}_k^{\mathrm{T}}} H(\boldsymbol{\psi}_{k+1}, \mathbf{x}_k, \mathbf{u}_k)\, \delta\mathbf{x}_k + H(\boldsymbol{\psi}_{k+1}, \mathbf{x}_k, \mathbf{u}_k + \delta\mathbf{u}_k) - H(\boldsymbol{\psi}_{k+1}, \mathbf{x}_k, \mathbf{u}_k) \right]$$
$$\simeq \sum_{k=0}^{N-1} \left[\frac{\partial}{\partial \mathbf{x}_k^{\mathrm{T}}} H(\boldsymbol{\psi}_{k+1}, \mathbf{x}_k, \mathbf{u}_k)\, \delta\mathbf{x}_k + \frac{\partial}{\partial \mathbf{u}_k^{\mathrm{T}}} H(\boldsymbol{\psi}_{k+1}, \mathbf{x}_k, \mathbf{u}_k)\, \delta\mathbf{u}_k \right]$$

Hence the expression of δJ:

$$\delta J \simeq \left. \frac{dh}{d\mathbf{x}_k^{\mathrm{T}}} \right|_{k=N} \cdot \delta\mathbf{x}_N + \boldsymbol{\psi}_N^{\mathrm{T}} \cdot \delta\mathbf{x}_N - \boldsymbol{\psi}_0^{\mathrm{T}} \cdot \delta\mathbf{x}_0$$
$$+ \sum_{k=0}^{N-1} \boldsymbol{\psi}_k^{\mathrm{T}} \cdot \delta\mathbf{x}_k - \sum_{k=0}^{N-1} \left[\frac{\partial}{\partial \mathbf{x}_k^{\mathrm{T}}} H(\boldsymbol{\psi}_{k+1}, \mathbf{x}_k, \mathbf{u}_k)\, \delta\mathbf{x}_k + \frac{\partial}{\partial \mathbf{u}_k^{\mathrm{T}}} H(\boldsymbol{\psi}_{k+1}, \mathbf{x}_k, \mathbf{u}_k)\, \delta\mathbf{u}_k \right],$$

or, by grouping on one hand the terms in $\delta\mathbf{x}_N$, on the other hand those in $\delta\mathbf{x}_k$:

$$\delta J \simeq \left[\left. \frac{dh}{d\mathbf{x}} \right|_{k=N} + \boldsymbol{\psi}_N \right]^{\mathrm{T}} \cdot \delta\mathbf{x}_N - \boldsymbol{\psi}_0^{\mathrm{T}} \cdot \delta\mathbf{x}_0$$
$$+ \sum_{k=0}^{N-1} \left[\boldsymbol{\psi}_k - \frac{\partial H(\boldsymbol{\psi}_{k+1}, \mathbf{x}_k, \mathbf{u}_k)}{\partial \mathbf{x}_k} \right]^{\mathrm{T}} \cdot \delta\mathbf{x}_k - \sum_{k=0}^{N-1} \frac{\partial H(\boldsymbol{\psi}_{k+1}, \mathbf{x}_k, \mathbf{u}_k)}{\partial \mathbf{u}_k^{\mathrm{T}}} \delta\mathbf{u}_k . \quad (3.67)$$

The objective being, as in the continuous case, to make $\delta J = 0$, the calculation goes on as follows:

1. one notices, as in the continuous case, that $\delta\mathbf{x}_0 = 0$;

2. $\boldsymbol{\psi}_k$ will be chosen as solution of the *adjoint equation*

$$\boldsymbol{\psi}_k = \frac{\partial}{\partial \mathbf{x}_k} H(\boldsymbol{\psi}_{k+1}, \mathbf{x}_k, \mathbf{u}_k) \ , \tag{3.68}$$

with the *final condition*, or *transversality condition*:

$$\boldsymbol{\psi}_N = -\left.\frac{d\,h}{d\,\mathbf{x}_k}\right|_{k=N} \ . \tag{3.69}$$

By taking into account (3.65) and (B.16), the adjoint equation is written more explicitly as a difference equation:

$$\boldsymbol{\psi}_k = -\frac{\partial f_0(\mathbf{x}_k, \mathbf{u}_k)}{\partial \mathbf{x}_k} + \frac{\partial \mathbf{f}^{\mathrm{T}}(\mathbf{x}_k, \mathbf{u}_k)}{\partial \mathbf{x}_k} \cdot \boldsymbol{\psi}_{k+1} \, . \tag{3.70}$$

3. Consequently, the equation $\delta J = 0$ is reduced to:

$$\delta J = -\sum_{k=0}^{N-1} \frac{\partial}{\partial\, \mathbf{u}_k^{\mathrm{T}}} H(\boldsymbol{\psi}_{k+1}, \mathbf{x}_k, \mathbf{u}_k)\, \delta \mathbf{u}_k = 0 \ .$$

3.6.3 Recapitulation of the Discrete Optimal Control

Suppose that $\mathbf{u}_k$ and $\mathbf{x}_k$, $k = 0,1,\ldots,N$, represent the optimal control law and the corresponding state trajectory, solutions of the optimal control problem defined by (3.61) and (3.62). There exists then an adjoint trajectory $\boldsymbol{\psi}_k$, $k = 0,1,\ldots,N$, such that the following relations hold, simultaneously:

$$\mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k), \ \ \mathbf{x}_0 \text{ given}$$

$$\begin{cases} \boldsymbol{\psi}_k = -\dfrac{\partial f_0(\mathbf{x}_k, \mathbf{u}_k)}{\partial \mathbf{x}_k} + \dfrac{\partial \mathbf{f}^{\mathrm{T}}(\mathbf{x}_k, \mathbf{u}_k)}{\partial \mathbf{x}_k} \cdot \boldsymbol{\psi}_{k+1} \\ \boldsymbol{\psi}_N = -\dfrac{d h(\mathbf{x}_N)}{d \mathbf{x}_k} \end{cases}$$

$$\frac{\partial}{\partial\, \mathbf{u}_k} H(\boldsymbol{\psi}, \mathbf{x}, \mathbf{u})\, \delta \mathbf{u}_k = 0 \ ,$$

where H is the Hamiltonian (3.64).

3.6.4 Case of the Discrete Linear Time-invariant Systems

The starting data of the problem are:

1. the discrete-time state model of a linear, time-invariant plant:

$$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}\mathbf{u}_k\,,\ \mathbf{x}_0 \ \text{ given}\,; \tag{3.71}$$

2. a quadratic cost functional:

$$J = \underbrace{\frac{1}{2}\mathbf{x}_N^{\mathrm{T}}\mathbf{S}\mathbf{x}_N}_{h(\mathbf{x}_N)} + \sum_{k=0}^{N-1}\underbrace{\frac{1}{2}(\mathbf{x}_k^{\mathrm{T}}\mathbf{Q}\,\mathbf{x}_k + \mathbf{u}_k^{\mathrm{T}}\mathbf{R}\,\mathbf{u}_k)}_{f_0(\mathbf{x}_k,\mathbf{u}_k)}\,, \tag{3.72}$$

where **Q**, **R** and **S** satisfy the same hypotheses as in the continuous case.

Since the final state is not submitted to constraint, the transversality condition yields:

$$\boldsymbol{\psi}_N = -\left.\frac{dh}{d\mathbf{x}_k}\right|_{k=N} = -\frac{d}{d\mathbf{x}_N}(\frac{1}{2}\mathbf{x}_N^{\mathrm{T}}\,\mathbf{S}\,\mathbf{x}_N) = -\,\mathbf{S}\mathbf{x}_N\,. \tag{3.73}$$

Furthermore, according to (3.65):

$$H(\boldsymbol{\psi}_{k+1},\mathbf{x}_k,\mathbf{u}_k) = -\frac{1}{2}(\mathbf{x}_k^{\mathrm{T}}\mathbf{Q}\,\mathbf{x}_k + \mathbf{u}_k^{\mathrm{T}}\mathbf{R}\,\mathbf{u}_k) + \boldsymbol{\psi}_{k+1}^{\mathrm{T}}\cdot(\mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}\mathbf{u}_k)\,.$$

Let us calculate $\partial H/\partial \mathbf{x}_k$. By successive use of (B.10) and (B.8), this expression becomes:

$$\frac{\partial H}{\partial \mathbf{x}_k} = -\,\mathbf{Q}\mathbf{x}_k + \mathbf{\Phi}^{\mathrm{T}}\boldsymbol{\psi}_{k+1}\,,$$

and the adjoint equation writes:

$$\boldsymbol{\psi}_k = -\,\mathbf{Q}\mathbf{x}_k + \mathbf{\Phi}^{\mathrm{T}}\boldsymbol{\psi}_{k+1}\,. \tag{3.74}$$

Since the value of **u** which maximizes *H* is given by

$$\frac{\partial H}{\partial \mathbf{u}_k} = 0\,,$$

the optimal control will be solution of

$$\frac{\partial H}{\partial \mathbf{u}_k} = -\mathbf{R}\,\mathbf{u}_k + \boldsymbol{\Gamma}^{\mathrm{T}} \boldsymbol{\psi}_{k+1} = 0 \,. \tag{3.75}$$

Since a state feedback control law is expected, we will try again to find $\boldsymbol{\psi}$ in the form

$$\boldsymbol{\psi}_k = -\,\mathbf{P}_k\,\mathbf{x}_k \,, \tag{3.76}$$

i.e.

$$\boldsymbol{\psi}_{k+1} = -\,\mathbf{P}_{k+1}\,\mathbf{x}_{k+1} = -\,\mathbf{P}_{k+1}(\boldsymbol{\Phi}\,\mathbf{x}_k + \boldsymbol{\Gamma}\,\mathbf{u}_k) \,.$$

Equation (3.75) giving the optimal control law writes then:

$$-\,\mathbf{R}\,\mathbf{u}_k - \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}(\boldsymbol{\Phi}\,\mathbf{x}_k + \boldsymbol{\Gamma}\,\mathbf{u}_k) = 0 \,,$$

$$-\,(\mathbf{R} + \boldsymbol{\Gamma}^{\mathrm{T}}\,\mathbf{P}_{k+1}\,\boldsymbol{\Gamma})\,\mathbf{u}_k = \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\,\boldsymbol{\Phi}\,\mathbf{x}_k \,.$$

Hence the optimal control law:

$$\mathbf{u}_k^* = -\,(\mathbf{R} + \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\,\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\,\boldsymbol{\Phi}\,\mathbf{x}_k \;\;, \tag{3.77}$$

which is of the desired state-feedback form:

$$\mathbf{u}_k^* = -\mathbf{L}_k\;\mathbf{x}_k \;\;, \tag{3.78}$$

with:

$$\mathbf{L}_k = (\mathbf{R} + \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\,\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\;\boldsymbol{\Phi} \;\;. \tag{3.79}$$

Rewriting the closed-loop state equation with this control law yields:

$$\mathbf{x}_{k+1} = \boldsymbol{\Phi}\,\mathbf{x}_k - \boldsymbol{\Gamma}\,\mathbf{L}_k\,\mathbf{x}_k = (\boldsymbol{\Phi} - \boldsymbol{\Gamma}\,\mathbf{L}_k)\,\mathbf{x}_k \,. \tag{3.80}$$

The substitution of (3.76) into (3.74) yields further:

$$-\,\mathbf{P}\,\mathbf{x}_k = -\,\mathbf{Q}\,\mathbf{x}_k - \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{P}_{k+1}\,\mathbf{x}_{k+1} \,,$$

which, by use of (3.80), writes now as follows:

$$\begin{aligned} -\mathbf{P}_k\,\mathbf{x}_k &= -\,\mathbf{Q}\,\mathbf{x}_k - \mathbf{\Phi}^{\mathrm{T}}\mathbf{P}_{k+1}\,(\mathbf{\Phi}-\mathbf{\Gamma}\,\mathbf{L}_k)\,\mathbf{x}_k \\ &= -\left[\mathbf{Q}+\mathbf{\Phi}^{\mathrm{T}}\mathbf{P}_{k+1}\,(\mathbf{\Phi}-\mathbf{\Gamma}\,\mathbf{L}_k)\right]\mathbf{x}_k \quad . \end{aligned}$$

Since this relation must be satisfied whatever $\mathbf{x}_k$, the following must be true:

$$\mathbf{P}_k = \mathbf{Q}+\mathbf{\Phi}^{\mathrm{T}}\mathbf{P}_{k+1}\,(\mathbf{\Phi}-\mathbf{\Gamma}\,\mathbf{L}_k)\,,$$

which yields after substitution of $\mathbf{L}_k$ by its value (3.79):

$$\mathbf{P}_k = \mathbf{\Phi}^{\mathrm{T}}\mathbf{P}_{k+1}\,\mathbf{\Phi} - \mathbf{\Phi}^{\mathrm{T}}\mathbf{P}_{k+1}\,\mathbf{\Gamma}(\mathbf{R}+\mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\,\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\,\mathbf{\Phi}+\mathbf{Q} \quad . \tag{3.81}$$

This equation is the *Riccati difference equation*. The suitable solution is a matrix $\mathbf{P}_k$, symmetric, positive semidefinite or positive definite according to the situation (see the discussion in Sect. 3.5.4).

The final condition (3.73) yields, as in the continuous case:

$$\mathbf{P}_N = \mathbf{S}\,. \tag{3.82}$$

Stationary solution: Assume, as in the continuous case, that $t_f \to \infty$, thus here that $N \to \infty$. The quadratic cost functional is now $J=\sum_{k=0}^{\infty}\frac{1}{2}(\mathbf{x}_k^{\mathrm{T}}\mathbf{Q}\,\mathbf{x}_k+\mathbf{u}_k^{\mathrm{T}}\mathbf{R}\,\mathbf{u}_k)$. The stationary equilibrium which takes then place obeys to

$$\mathbf{P}_{k+1} = \mathbf{P}_k = \mathbf{P} = \text{constant}\,,$$

and the Riccati difference equation is replaced by:

$$\mathbf{P} = \mathbf{\Phi}^{\mathrm{T}}\mathbf{P}\mathbf{\Phi} - \mathbf{\Phi}^{\mathrm{T}}\mathbf{P}\mathbf{\Gamma}(\mathbf{R}+\mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}\mathbf{\Phi}+\mathbf{Q} \quad . \tag{3.83}$$

This is the *algebraic Riccati equation* of the *discrete time systems*, in short: *discrete algebraic Riccati equation* (DARE).

P is, as previously, a symmetric matrix. Here again the suitable solution is the unique matrix **P** which is positive semidefinite or positive definite, according to the situation (for the existence conditions of this solution, see Sect. 3.5.6).

The optimal state feedback matrix becomes then the constant matrix:

$$\mathbf{L} = (\mathbf{R}+\mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}\mathbf{\Phi} \quad . \tag{3.84}$$

3.7 Optimal Control Robustness

3.7.1 Continuous-time Systems

3.7.1.1 Return Difference

Let us redraw the closed-loop system equipped with a state-feedback control law $\mathbf{u} = -\mathbf{L}\,\mathbf{x}$ as a unity feedback system, as illustrated in Fig. 3.2.

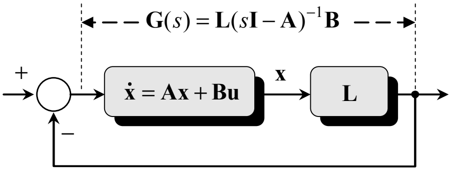

$\mathbf{G}(s) = \mathbf{L}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$
+
−
$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$
x
L

Fig. 3.2 State-feedback system represented as a unity feedback loop.

Let us start then with the algebraic Riccati equation (3.54) which corresponds to the stationary behavior:

$$\mathbf{P}\mathbf{A} + \mathbf{A}^{\mathrm{T}}\mathbf{P} - \mathbf{P}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P} + \mathbf{Q} = \mathbf{0}\ ,$$

and add to it $+s\mathbf{P}$ and $-s\mathbf{P}$ after having changed the sign of the first side [AnMo89]:

$$\mathbf{P}(s\mathbf{I} - \mathbf{A}) + (-s\mathbf{I} - \mathbf{A}^{\mathrm{T}})\mathbf{P} + \mathbf{P}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P} = \mathbf{Q}\,.$$

Multiply this equation left by $\mathbf{B}^{\mathrm{T}}(-s\mathbf{I} - \mathbf{A}^{\mathrm{T}})^{-1}$, and right by $(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$:

$$\begin{aligned}\mathbf{B}^{\mathrm{T}}(-s\mathbf{I} - \mathbf{A}^{\mathrm{T}})^{-1}\mathbf{P}\mathbf{B} + \mathbf{B}^{\mathrm{T}}\mathbf{P}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{B}^{\mathrm{T}}(-s\mathbf{I} - \mathbf{A}^{\mathrm{T}})^{-1}\mathbf{P}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\\ = \mathbf{B}^{\mathrm{T}}(-s\mathbf{I} - \mathbf{A}^{\mathrm{T}})^{-1}\mathbf{Q}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\,.\end{aligned}$$

If we remark that, according to (3.55), $\mathbf{R}\,\mathbf{L} = \mathbf{B}^{\mathrm{T}}\mathbf{P}$, this equation rewrites

$$\begin{aligned}\mathbf{B}^{\mathrm{T}}(-s\mathbf{I} - \mathbf{A}^{\mathrm{T}})^{-1}\mathbf{L}^{\mathrm{T}}\mathbf{R} + \mathbf{R}\,\mathbf{L}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{B}^{\mathrm{T}}(-s\mathbf{I} - \mathbf{A}^{\mathrm{T}})^{-1}\mathbf{L}^{\mathrm{T}}\mathbf{R}\,\mathbf{L}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\\ = \mathbf{B}^{\mathrm{T}}(-s\mathbf{I} - \mathbf{A}^{\mathrm{T}})^{-1}\mathbf{Q}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\,.\end{aligned}$$

Remarking furthermore that the first side can also be written

$$\left[\mathbf{I}+\mathbf{B}^{\mathrm{T}}(-s\mathbf{I}-\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{L}^{\mathrm{T}}\right]\mathbf{R}\left[\mathbf{I}+\mathbf{L}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\right]-\mathbf{R}\,,$$

we obtain finally the equality

$$\begin{aligned}&\left[\mathbf{I}+\mathbf{L}(-s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\right]^{\mathrm{T}}\mathbf{R}\left[\mathbf{I}+\mathbf{L}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\right]\\&\qquad=\mathbf{R}+\mathbf{B}^{\mathrm{T}}(-s\mathbf{I}-\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{Q}\,(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\;,\end{aligned}\tag{3.85}$$

called *return difference equality* [AnMo89], from the name *return difference* which has been given to the quantity $\mathbf{I}+\mathbf{L}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}$, when the plant and the controller are organized as shown in Fig. 3.2, and which reminds of the term $1+G$ encountered in the scalar case.

3.7.1.2 Frequency Behavior (Nyquist Diagram)

Let $s=j\omega$. Eq. (3.85) becomes:

$$\begin{aligned}&\left[\mathbf{I}+\mathbf{L}(-j\omega\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\right]^{\mathrm{T}}\mathbf{R}\left[\mathbf{I}+\mathbf{L}(j\omega\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\right]\\&\qquad=\mathbf{R}+\mathbf{B}^{\mathrm{T}}(-j\omega\mathbf{I}-\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{Q}\,(j\omega\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\;.\end{aligned}$$

In the SISO case, with the substitutions $\mathbf{B}=\mathbf{b}$, $\mathbf{R}=r$ and $\mathbf{L}=\boldsymbol{\ell}^{\mathrm{T}}$, this expression becomes:

$$r\left|1+\boldsymbol{\ell}^{\mathrm{T}}(j\omega\mathbf{I}-\mathbf{A})^{-1}\mathbf{b}\right|^{2}=r+\mathbf{b}^{\mathrm{T}}(-j\omega-\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{Q}\,(j\omega-\mathbf{A})^{-1}\mathbf{b}\,.$$

$\mathbf{Q}$ being symmetric, we can write $\mathbf{Q}=\mathbf{q}\mathbf{q}^{\mathrm{T}}$, and the previous expression becomes

$$r\left|1+\boldsymbol{\ell}^{\mathrm{T}}(j\omega\mathbf{I}-\mathbf{A})^{-1}\mathbf{b}\right|^{2}=r+\left|\mathbf{q}^{\mathrm{T}}(j\omega-\mathbf{A})^{-1}\mathbf{b}\right|^{2}.$$

Since $r>0$, this yields

$$\left|1+\boldsymbol{\ell}^{\mathrm{T}}(j\omega\mathbf{I}-\mathbf{A})^{-1}\mathbf{b}\right|^{2}>1,$$

or

$$\left|1+\boldsymbol{\ell}^{\mathrm{T}}(j\omega\mathbf{I}-\mathbf{A})^{-1}\mathbf{b}\right|>1.\tag{3.86}$$

In order to draw conclusions on the closed-loop system stability from this inequality, we must plot the open-loop Nyquist diagram.

In the SISO case, the open-loop transfer matrix $\mathbf{G}(s)$, indicated in Fig. 3.2, becomes the following transfer function:

$$\mathbf{G}(s) = G(s) = \boldsymbol{\ell}^{\mathrm{T}}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}\,.$$

The open-loop frequency response is thus given by:

$$G(j\omega) = \boldsymbol{\ell}^{\mathrm{T}}(j\omega\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}\,.$$

Inequality (3.86) shows then that the distance of any point of the Nyquist diagram to the critical point -1 will always be greater than, or equal to, 1. In other words, the Nyquist diagram will never enter into the circle of unit radius, centered at $(-1+j0)$, as illustrated in Fig. 3.3.

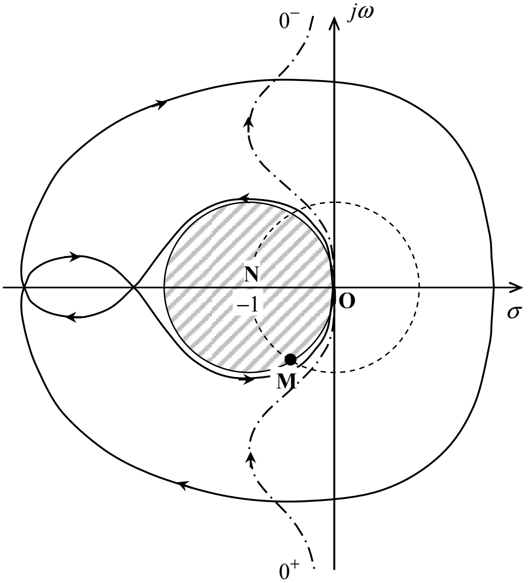


Fig. 3.3 Example of two fictitious Nyquist plots of $\ell^{\mathrm{T}}(j\omega\mathbf{I}-\mathbf{A})^{-1}\mathbf{b}$ (solid line and dash-dotted line) avoiding the critical unit disc (hatched).

If M denotes the intersection point of this circle with the unit radius circle centered at O, it appears that the triangle OMN is equilateral. Thus, the smallest possible phase margin amounts to 60° and the gain margin is infinite. In the opposite direction, the closed-loop stability is guaranteed as long as the open-loop gain is not multiplied by less than $1/2$.

3.7.2 Discrete-time Systems

3.7.2.1 Return Difference

Let us start here with the discrete algebraic Riccati equation, (3.83):

$$\mathbf{P}=\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Phi}-\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Gamma}(\mathbf{R}+\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Phi}+\mathbf{Q}\ ,$$

or also

$$\mathbf{P}-\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Phi}+\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Gamma}(\mathbf{R}+\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Phi}=\mathbf{Q}\,. \tag{3.87}$$

Rewrite now the first two terms of the left side, by adding and subtracting to them the quantity $\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{P}\,z$:

$$\begin{aligned}\mathbf{P}-\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Phi}= &\ \underbrace{\mathbf{P}-\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{P}z}\ +\ \underbrace{\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{P}z-\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Phi}}\\ &=(z^{-1}\mathbf{I}-\boldsymbol{\Phi}^{\mathrm{T}})\mathbf{P}z\ +\ \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{P}(z\mathbf{I}-\boldsymbol{\Phi})\ ,\end{aligned}$$

then, by adding and subtracting the quantity $(z^{-1}\mathbf{I}-\boldsymbol{\Phi}^{\mathrm{T}})\,\mathbf{P}\,\boldsymbol{\Phi}$:

$$\begin{aligned}&\mathbf{P}-\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Phi}\\ &\quad=\underbrace{(z^{-1}\mathbf{I}-\boldsymbol{\Phi}^{\mathrm{T}})\mathbf{P}z-(z^{-1}\mathbf{I}-\boldsymbol{\Phi}^{\mathrm{T}})\,\mathbf{P}\,\boldsymbol{\Phi}}+(z^{-1}\mathbf{I}-\boldsymbol{\Phi}^{\mathrm{T}})\,\mathbf{P}\,\boldsymbol{\Phi}+\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{P}(z\mathbf{I}-\boldsymbol{\Phi})\\ &\quad=\qquad(z^{-1}\mathbf{I}-\boldsymbol{\Phi}^{\mathrm{T}})\mathbf{P}(z\mathbf{I}-\boldsymbol{\Phi})\qquad+(z^{-1}\mathbf{I}-\boldsymbol{\Phi}^{\mathrm{T}})\,\mathbf{P}\,\boldsymbol{\Phi}+\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{P}(z\mathbf{I}-\boldsymbol{\Phi})\,.\end{aligned} \tag{3.88}$$

Multiply now the two sides of (3.87) by $\boldsymbol{\Gamma}^{\mathrm{T}}(z^{-1}\mathbf{I}-\boldsymbol{\Phi}^{\mathrm{T}})^{-1}$ to the left, and by $(z\mathbf{I}-\boldsymbol{\Phi})^{-1}\boldsymbol{\Gamma}$ to the right, while using (3.88):

$$\begin{aligned}&\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Gamma}+\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Phi}(z\mathbf{I}-\boldsymbol{\Phi})^{-1}\boldsymbol{\Gamma}+\boldsymbol{\Gamma}^{\mathrm{T}}(z^{-1}\mathbf{I}-\boldsymbol{\Phi}^{\mathrm{T}})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Gamma}\\ &\quad+\boldsymbol{\Gamma}^{\mathrm{T}}(z^{-1}\mathbf{I}-\boldsymbol{\Phi}^{\mathrm{T}})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Gamma}\left(\mathbf{R}+\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Phi}(z\mathbf{I}-\boldsymbol{\Phi})^{-1}\boldsymbol{\Gamma}\\ &\qquad\qquad=\boldsymbol{\Gamma}^{\mathrm{T}}(z^{-1}\mathbf{I}-\boldsymbol{\Phi}^{\mathrm{T}})^{-1}\mathbf{Q}(z\mathbf{I}-\boldsymbol{\Phi})^{-1}\boldsymbol{\Gamma}\,.\end{aligned} \tag{3.89}$$

If we use now (3.84), which can be rewritten as

$$(\mathbf{R}+\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Gamma})\,\mathbf{L}=\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}\,\boldsymbol{\Phi}, \tag{3.90}$$

(3.89) can be put in the following form:

$$\underbrace{\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Gamma}}_{\textcircled{1}}+\underbrace{(\mathbf{R}+\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Gamma})\mathbf{L}(z\mathbf{I}-\boldsymbol{\Phi})^{-1}\boldsymbol{\Gamma}}_{\textcircled{2}}+\underbrace{\boldsymbol{\Gamma}^{\mathrm{T}}(z^{-1}\mathbf{I}-\boldsymbol{\Phi}^{\mathrm{T}})^{-1}\mathbf{L}^{\mathrm{T}}(\mathbf{R}+\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Gamma})}_{\textcircled{3}}$$
$$+\underbrace{\boldsymbol{\Gamma}^{\mathrm{T}}(z^{-1}\mathbf{I}-\boldsymbol{\Phi}^{\mathrm{T}})^{-1}\mathbf{L}^{\mathrm{T}}(\mathbf{R}+\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Gamma})\mathbf{L}(z\mathbf{I}-\boldsymbol{\Phi})^{-1}\boldsymbol{\Gamma}}_{\textcircled{4}}$$
$$=\boldsymbol{\Gamma}^{\mathrm{T}}(z^{-1}\mathbf{I}-\boldsymbol{\Phi}^{\mathrm{T}})^{-1}\mathbf{Q}(z\mathbf{I}-\boldsymbol{\Phi})^{-1}\boldsymbol{\Gamma},$$

or also, after addition of **R** to the two sides:

$$\Big[\overbrace{\mathbf{I}+}^{a}\overbrace{\mathbf{L}(z^{-1}\mathbf{I}-\boldsymbol{\Phi})^{-1}\boldsymbol{\Gamma}}^{b}\Big]^{\mathrm{T}}\overbrace{(\mathbf{R}+\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Gamma})}^{c}\Big[\overbrace{\mathbf{I}+}^{d}\overbrace{\mathbf{L}(z\mathbf{I}-\boldsymbol{\Phi})^{-1}\boldsymbol{\Gamma}}^{e}\Big]$$
$$=\mathbf{R}+\boldsymbol{\Gamma}^{\mathrm{T}}(z^{-1}\mathbf{I}-\boldsymbol{\Phi}^{\mathrm{T}})^{-1}\mathbf{Q}(z\mathbf{I}-\boldsymbol{\Phi})^{-1}\boldsymbol{\Gamma}\,. \qquad (3.91)$$

The identity of these two formulae is easily proved by verifying that:

$$a\,c\,d=\mathbf{R}+\textcircled{1} \qquad a\,c\,e=\textcircled{2}$$
$$b^{\mathrm{T}}c\,d=\textcircled{3} \qquad b^{\mathrm{T}}c\,e=\textcircled{4}$$

(3.91) is very similar to (3.85) of the continuous case, where in addition to the conventional replacement of **A** by $\boldsymbol{\Phi}$ and of **B** by $\boldsymbol{\Gamma}$, one would have replaced $-s$ by z^{-1} and **R** by $(\mathbf{R}+\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Gamma})$.

3.7.2.2 Frequency Response (Nyquist Diagram)

Let $z=e^{j\omega T_s}$, as is usual to study the frequency behavior of a discrete-time system, resulting from the sampling at a period T_s of a continuous-time system. (3.91) becomes:

$$\Big[\mathbf{I}+\mathbf{L}(e^{-j\omega T_s}\,\mathbf{I}-\boldsymbol{\Phi})^{-1}\boldsymbol{\Gamma}\Big]^{\mathrm{T}}(\mathbf{R}+\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Gamma})\Big[\mathbf{I}+\mathbf{L}(e^{j\omega T_s}\,\mathbf{I}-\boldsymbol{\Phi})^{-1}\boldsymbol{\Gamma}\Big]$$
$$=\mathbf{R}+\boldsymbol{\Gamma}^{\mathrm{T}}(e^{-j\omega T_s}\,\mathbf{I}-\boldsymbol{\Phi}^{\mathrm{T}})^{-1}\mathbf{Q}(e^{j\omega T_s}\,\mathbf{I}-\boldsymbol{\Phi})^{-1}\boldsymbol{\Gamma}\,.$$

In the SISO case, with the substitutions $\boldsymbol{\Gamma}=\boldsymbol{\gamma}$, $\mathbf{R}=r$ and $\mathbf{L}=\boldsymbol{\ell}^{\mathrm{T}}$, this expression writes

$$(r+\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{P}\boldsymbol{\gamma})\left|1+\boldsymbol{\ell}^{\mathrm{T}}(e^{j\omega T_s}\,\mathbf{I}-\boldsymbol{\Phi})^{-1}\boldsymbol{\gamma}\right|^2=r+\boldsymbol{\gamma}^{\mathrm{T}}(e^{-j\omega T_s}\,\mathbf{I}-\boldsymbol{\Phi}^{\mathrm{T}})^{-1}\mathbf{Q}(e^{j\omega T_s}\,\mathbf{I}-\boldsymbol{\Phi})^{-1}\boldsymbol{\gamma}$$

and, by letting as in the continuous case $\mathbf{Q}=\mathbf{q}\mathbf{q}^{\mathrm{T}}$, since **Q** is symmetric,

$$(r+\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{P}\boldsymbol{\gamma})\left|1+\boldsymbol{\ell}^{\mathrm{T}}(e^{j\omega T_s}\,\mathbf{I}-\boldsymbol{\Phi})^{-1}\boldsymbol{\gamma}\right|^2 = r+\left|\mathbf{q}^{\mathrm{T}}(e^{j\omega T_s}\,\mathbf{I}-\boldsymbol{\Phi})^{-1}\boldsymbol{\gamma}\right|^2,$$

which yields

$$(r+\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{P}\boldsymbol{\gamma})\left|1+\boldsymbol{\ell}^{\mathrm{T}}(e^{j\omega T_s}\,\mathbf{I}-\boldsymbol{\Phi})^{-1}\boldsymbol{\gamma}\right|^2 \geq r\,,$$

or also, since $r>0$,

$$\left|1+\boldsymbol{\ell}^{\mathrm{T}}(e^{j\omega T_s}\,\mathbf{I}-\boldsymbol{\Phi})^{-1}\boldsymbol{\gamma}\right| \geq \rho\,,$$

where:

$$\rho = \frac{1}{\sqrt{1+r^{-1}\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{P}\,\boldsymbol{\gamma}}} < 1\ . \qquad (3.92)$$

The Nyquist plot remains therefore out of a disk of radius ρ centered at -1 (Fig. 3.4). This has following consequences:

1. for the gain margin G_M:

$$\begin{aligned}\lim\nolimits_{\inf} G_M &= (1+\rho)^{-1}\\ \lim\nolimits_{\sup} G_M &= (1-\rho)^{-1}\end{aligned}$$

2. for the phase margin Φ_M:

$$\sin(\frac{1}{2}\Phi_M) = \frac{\rho}{2},\ \Rightarrow\ \Phi_M = 2\arcsin\frac{\rho}{2}\,.$$

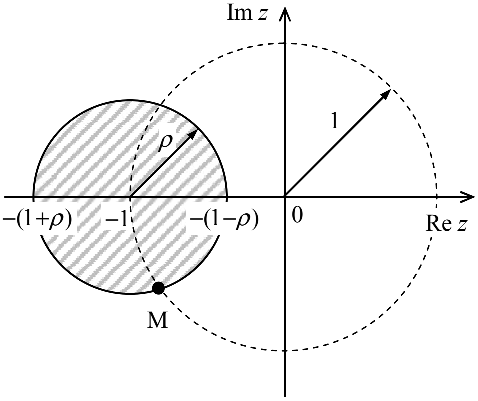


Fig. 3.4 Disk avoided by any Nyquist plot for discrete-time optimal control (open-loop).

3.8 Dynamic Programming. Relation to the Principle of the Maximum

3.8.1 Optimality Principle

Assume that a given optimal control problem, defined on a time interval $[0, t_f]$, has been solved, and call it Problem I: the solution to this problem produces some optimal trajectory.

Let us define now the following Problem II: given an instant $t \in [0, t_f]$ and the corresponding point $\mathbf{x}(t)$ of the optimal trajectory determined above, find the solution of the new problem of optimal control starting from this point, which is taken as new initial state.

The optimality principle claims that the solution to Problem II corresponds exactly to the rest of the solution to Problem I, from this point on. Let us reformulate this principle, stated by Bellman:

Optimality Principle. *Starting from any arbitrary point of an optimal trajectory, the remaining part of the trajectory is optimal for the optimization problem corresponding to this point taken as new initial state.*

3.8.2 Continuous-time Systems. Hamilton-Jacobi-Bellman Equation

Hypotheses:

$$\dot{\mathbf{x}} = \mathbf{f}\left[\mathbf{x}(t), \mathbf{u}(t)\right],$$

$$\mathbf{x}(0) = \mathbf{x}_0 \text{ given},$$

$$\mathbf{u}(t) \in \mathscr{U},$$

$$J = h\left[\mathbf{x}(t_f)\right] + \int_0^{t_f} f_0\left[\mathbf{x}(t), \mathbf{u}(t)\right] dt .$$

Let $V\left[\mathbf{x}(t), t\right]$ denote the optimal (minimum) cost functional associated with the motion from some state $\mathbf{x}(t)$ until the final state $\mathbf{x}(t_f)$:

$$V\left[\mathbf{x}(t),t\right]=V\left[\mathbf{x},t\right]$$
$$=\min_{\mathbf{u}}\left\{h\left[\mathbf{x}(t_f)\right]+\int_t^{t_f} f_0\left[\mathbf{x}(\tau),\mathbf{u}(\tau)\right]d\tau\right\}. \qquad (3.93)$$

The present goal is to find an equation for $V\left[\mathbf{x},t\right]$.

We start with the hypothesis that V is known for $t+\delta t$, where δt is a small time interval, and proceed backwards towards time t. By application of the optimality principle,

$$V(\mathbf{x},t)=\min_{\mathbf{u}}\left\{\int_t^{t+\delta t} f_0\left[\mathbf{x}(\tau),\mathbf{u}(\tau)\right]d\tau+V(\mathbf{x}+\delta\mathbf{x},t+\delta t)\right\}, \qquad (3.94)$$

where the integration interval $[t,t_f]$ has been split into $[t,t+\delta t]$ and $[t+\delta t,t_f]$. If we suppose that during the first interval the applied control, $\mathbf{u}(\tau)$, remains practically constant, we can write:

$$\int_t^{t+\delta t} f_0\left[\mathbf{x}(\tau),\mathbf{u}(\tau)\right]d\tau \simeq f_0\left[\mathbf{x}(t),\mathbf{u}(t)\right]\delta t\,.$$

Moreover, by first order Taylor series approximation,

$$V(\mathbf{x}+\delta\mathbf{x},t+\delta t)\simeq V(\mathbf{x},t)+\frac{dV(\mathbf{x},t)}{dt}\delta t$$
$$\simeq V(\mathbf{x},t)+\left[\frac{\partial V(\mathbf{x},t)}{\partial t}+\frac{\partial V(\mathbf{x},t)}{\partial \mathbf{x}^{\mathrm{T}}}\cdot\frac{d\mathbf{x}}{dt}\right]\delta t\,.$$

By substitution in (3.94), this yields:

$$V(\mathbf{x},t)=\min_{\mathbf{u}}\left\{f_0(\mathbf{x},\mathbf{u})\delta t+V(\mathbf{x},t)+\frac{\partial V(\mathbf{x},t)}{\partial t}\delta t+\frac{\partial V(\mathbf{x},t)}{\partial \mathbf{x}^{\mathrm{T}}}\cdot\mathbf{f}(\mathbf{x},\mathbf{u})\delta t\right\}.$$

Since $V(\mathbf{x},t)$ does not depend on $\mathbf{u}$, this term can be extracted from the second member brace, and then subtracted from both sides:

$$0=\min_{\mathbf{u}}\left\{f_0(\mathbf{x},\mathbf{u})\delta t+\frac{\partial V(\mathbf{x},t)}{\partial t}\delta t+\frac{\partial V(\mathbf{x},t)}{\partial \mathbf{x}^{\mathrm{T}}}\cdot\mathbf{f}(\mathbf{x},\mathbf{u})\delta t\right\}.$$

The second term of the sum in the brace does not depend either on $\mathbf{u}$, and can thus also be withdrawn from the minimization. By dividing further the two sides by δt, we obtain finally:

$$0=\frac{\partial V(\mathbf{x},t)}{\partial t}+\min_{\mathbf{u}}\left\{f_0(\mathbf{x},\mathbf{u})+\frac{\partial V(\mathbf{x},t)}{\partial \mathbf{x}^{\mathrm{T}}}\cdot\mathbf{f}(\mathbf{x},\mathbf{u})\right\} \ . \tag{3.95}$$

This equation is called *Hamilton-Jacobi-Bellman equation*. Its numerical solution is calculated backwards, by starting with the evident final condition according to (3.93):

$$V(\mathbf{x},t_f)=h[\mathbf{x}(t_f)] \ . \tag{3.96}$$

3.8.3 Relation to the Maximum Principle

Letting the costate vector of the previous method be

$$\boldsymbol{\psi}=-\frac{\partial V(\mathbf{x},t)}{\partial \mathbf{x}} \ , \tag{3.97}$$

the Hamiltonian of (3.4) is retrieved in the form

$$H=-f_0(\mathbf{x},\mathbf{u})-\frac{\partial V(\mathbf{x},t)}{\partial \mathbf{x}^{\mathrm{T}}}\cdot\mathbf{f}(\mathbf{x},\mathbf{u}) \ , \tag{3.98}$$

and (3.95) can be written:

$$0=\frac{\partial V(\mathbf{x},t)}{\partial t}+\min_{\mathbf{u}}\left[-H\right].$$

Remarking that $\min_{\mathbf{u}}\left[-H\right]=-\max_{\mathbf{u}}\left[H\right]$, we obtain the following equation:

$$0=\frac{\partial V(\mathbf{x},t)}{\partial t}-\max_{\mathbf{u}}\left[H\right]. \tag{3.99}$$

The Pontryagin's maximum principle has been regained here. The optimal control **u** is the one which maximizes the Hamiltonian.

3.8.4 Example: the LQC Problem

Consider the linear system $\dot{\mathbf{x}} = \mathbf{A}\,\mathbf{x} + \mathbf{B}\,\mathbf{u}$, and the quadratic criterion $J = \frac{1}{2}\int_0^{t_f} (\mathbf{x}^{\mathrm{T}}\mathbf{Q}\,\mathbf{x} + \mathbf{u}^{\mathrm{T}}\mathbf{R}\,\mathbf{u})\,dt$, where the simplifying assumption $\mathbf{S} = 0$ has been made.

Let us look for a solution of the Hamilton-Jacobi-Bellman equation of the form

$$V(\mathbf{x},t) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{P}(t)\,\mathbf{x}\,, \tag{3.100}$$

where $\mathbf{P}(t)$ is an $(n \times n)$, symmetric, unknown matrix.

The *partial* derivative of $V(\mathbf{x},t)$ with respect to **x**, thus at *t* constant and, hence, at $\mathbf{P}(t)$ constant, amounts, according to (B.10), to

$$\frac{\partial V(\mathbf{x},t)}{\partial \mathbf{x}} = \mathbf{P}(t)\,\mathbf{x}\,,$$

from where it results by transposition that

$$\frac{\partial V(\mathbf{x},t)}{\partial \mathbf{x}^{\mathrm{T}}} = \mathbf{x}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}(t) = \mathbf{x}^{\mathrm{T}}\mathbf{P}(t)\,,$$

so that the Hamiltonian of this system writes, according to (3.98):

$$H = -\frac{1}{2}(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\,\mathbf{x} + \mathbf{u}^{\mathrm{T}}\mathbf{R}\,\mathbf{u}) - \mathbf{x}^{\mathrm{T}}\mathbf{P}(t)\,(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u})\,.$$

Let us determine for which value of **u** this expression is maximum. According to (B.10) and (B.8),

$$\frac{\partial H}{\partial \mathbf{u}} = -\mathbf{R}\mathbf{u} - \left[\mathbf{x}^{\mathrm{T}}\mathbf{P}(t)\mathbf{B}\right]^{\mathrm{T}} = -\mathbf{R}\mathbf{u} - \mathbf{B}^{\mathrm{T}}\mathbf{P}(t)\mathbf{x}\,.$$

The maximum of *H* will thus be obtained for

$$\mathbf{u} = -\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}(t)\,\mathbf{x}\,,$$

and will thus have the following value:

$$H_{\max} = -\frac{1}{2}\left[\mathbf{x}^{\mathrm{T}}\mathbf{Q}\,\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{P}(t)\mathbf{B}(\mathbf{R}^{-1})^{\mathrm{T}}\mathbf{R}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}(t)\mathbf{x}\right] - \mathbf{x}^{\mathrm{T}}\mathbf{P}(t)\left[\mathbf{A}\mathbf{x} - \mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}(t)\mathbf{x}\right].$$

R being symmetric, $(\mathbf{R}^{-1})^{\mathrm{T}} = \mathbf{R}^{-1}$, and

$$H_{\max} = -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{Q}\,\mathbf{x} + \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{P}(t)\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}(t)\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{P}(t)\mathbf{A}\mathbf{x},$$

or, with

$$\mathbf{x}^{\mathrm{T}}\mathbf{P}(t)\mathbf{A}\mathbf{x} = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\left[\mathbf{P}(t)\mathbf{A} + \mathbf{A}^{\mathrm{T}}\mathbf{P}(t)\right]\mathbf{x},$$

$$H_{\max} = -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\left[\mathbf{Q} - \mathbf{P}(t)\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}(t) + \mathbf{P}(t)\mathbf{A} + \mathbf{A}^{\mathrm{T}}\mathbf{P}(t)\right]\mathbf{x}.$$

Equation (3.99) becomes thus:

$$0 = -\frac{\partial V(\mathbf{x},t)}{\partial t} - \frac{1}{2}\mathbf{x}^{\mathrm{T}}\left[\mathbf{Q} - \mathbf{P}(t)\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}(t) + \mathbf{P}(t)\mathbf{A} + \mathbf{A}^{\mathrm{T}}\mathbf{P}(t)\right]\mathbf{x}.$$

The *partial* derivative of $V(\mathbf{x},t)$ with respect to *t*, thus at $\mathbf{x}$ constant, writes:

$$\frac{\partial V(\mathbf{x},t)}{\partial t} = \frac{\partial}{\partial t}\left[\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{P}(t)\mathbf{x}\right] = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\,\dot{\mathbf{P}}(t)\,\mathbf{x}.$$

The Hamilton-Jacobi-Bellman equation is then:

$$0 = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\left[\dot{\mathbf{P}}(t) + \mathbf{Q} + \mathbf{P}(t)\mathbf{A} + \mathbf{A}^{\mathrm{T}}\mathbf{P}(t) - \mathbf{P}(t)\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}(t)\right]\mathbf{x}.$$

For that expression to be identically true, i.e. whatever *x*, $\mathbf{P}(t)$ must be chosen solution of the equation

$$-\dot{\mathbf{P}}(t) = \mathbf{P}(t)\mathbf{A} + \mathbf{A}^{\mathrm{T}}\mathbf{P}(t) - \mathbf{P}(t)\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}(t) + \mathbf{Q},$$

which is nothing else than the Riccati differential equation (3.35) found in Sect. 3.5.2.

Remark 3.13. Since, according to (3.93), $J_{\min} = V(\mathbf{x},0)$, we deduce from what precedes that

$$J_{\min} = \frac{1}{2}\mathbf{x}^{\mathrm{T}}(0)\,\mathbf{P}(0)\,\mathbf{x}(0) \ . \tag{3.101}$$

We have thus regained also the expression (3.43) of the cost functional corresponding to the optimal control law of Problem I.

3.9 Solved Exercises

Exercise 3.1 Optimization of a Mechanical System

Consider the continuous-time linear system described by the state equations

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = -\dfrac{1}{\tau_m} x_2 + \dfrac{K}{\tau_m} u \end{cases}$$

These equations describe the operation of a pair of scissors aimed at cutting on the fly a paper band moving at constant velocity. The state variables $x_1(t)$ and $x_2(t)$ are respectively the circular motion and the tangential speed of one of the two scissor blades, K and τ_m are the proportionality constant between speed and armature voltage of the motor driving the blade and its mechanical time constant. Since the cost of motors increases with their power, the goal of the present exercise consists in optimizing the control of this device so that the mean effective torque of the motor is minimized between two consecutive cuts, i.e. between the times 0 and t_f, which will determine the minimum motor power required. This amounts to minimizing a term proportional to the integral of the square of the blade angular acceleration, or tangential acceleration, on the considered time interval. The resulting minimization criterion will thus be

$$J = \frac{1}{2}\int_{t_0}^{t_f} \dot{x}_2^2 \, dt = \frac{1}{2}\int_{t_0}^{t_f} \left(-\frac{1}{\tau_m} x_2 + \frac{K}{\tau_m} u \right)^2 dt \, .$$

The final state $x(t_f)$ is of course entirely determined by the necessity for the two blades to be aligned against each other at the time t_f of the next cut and for their tangential speed to match then that of the linear motion of the paper band. To simplify the problem, it is not asked to introduce these requirements in the calculation.

Although the previous criterion is quadratic, it is not the usual LQC design criterion and the solving of this problem will not rely on the Riccati equation. Adopt instead the following approach:

a) Write the Hamiltonian expression and deduce from it the adjoint state equation and the optimal control law, expressed as a function of the costate.

b) Substitute this law into the state equation and the adjoint state equation.

c) Calculate the general solution of the obtained differential equations, by introducing as many arbitrary constants as required ($c_1, c_2, \dots$), and deduce from this the form of the optimal control law expressed as a function of these constants and of the model parameters.

Solution:

(a) With the notations of Sect. 3.2, equation (3.2) and followings, it holds that

$$f_0(\mathbf{x},u,t) = \frac{1}{2}\left(-\frac{1}{\tau_m}x_2 + \frac{K}{\tau_m}u\right)^2,$$

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x},u,t) \quad \Rightarrow \quad \mathbf{f} = \begin{pmatrix} x_2 \\ -\dfrac{1}{\tau_m}x_2 + \dfrac{K}{\tau_m}u \end{pmatrix}.$$

According to (3.4), with $\boldsymbol{\psi}^{\mathrm{T}} = \begin{pmatrix}\psi_1 & \psi_2\end{pmatrix}$, the Hamiltonian is thus

$$H = -\frac{1}{2}\left(-\frac{1}{\tau_m}x_2 + \frac{K}{\tau_m}u\right)^2 + \psi_1 x_2 + \psi_2\left(-\frac{1}{\tau_m}x_2 + \frac{K}{\tau_m}u\right),$$

which yields the adjoint state equation (3.13), componentwise:

$$\dot{\psi}_1 = -\frac{\partial H}{\partial x_1} = 0,$$

$$\dot{\psi}_2 = -\frac{\partial H}{\partial x_2} = -\left[\frac{1}{\tau_m}\left(-\frac{1}{\tau_m}x_2 + \frac{K}{\tau_m}u\right) + \psi_1 - \frac{1}{\tau_m}\psi_2\right].$$

To obtain the optimal control law u^*, maximize H with respect to u:

$$\frac{\partial H}{\partial u}=-\frac{K}{\tau_m}\left(-\frac{1}{\tau_m}x_2+\frac{K}{\tau_m}u^*\right)+\frac{K}{\tau_m}\psi_2=0\,,$$

$$u^*=\frac{1}{K}x_2+\frac{\tau_m}{K}\psi_2\,.$$

(b) By substitution,

$$\begin{aligned}
&\dot{x}_1=x_2\\
&\dot{x}_2=-\frac{1}{\tau_m}x_2+\frac{K}{\tau_m}u^*=-\frac{1}{\tau_m}x_2+\frac{1}{\tau_m}x_2+\psi_2=\psi_2\\
&\dot{\psi}_1=0\\
&\dot{\psi}_2=-\left[\frac{1}{\tau_m}\left(-\frac{1}{\tau_m}x_2+\frac{1}{\tau_m}x_2+\psi_2\right)+\psi_1-\frac{1}{\tau_m}\psi_2\right]=-\psi_1
\end{aligned}$$

(c) The general solution of these differential equations is therefore:

$$\begin{aligned}
&\psi_1=c_1\ \text{(constant)}\\
&\psi_2=-\int\psi_1(\tau)d\tau=-c_1t+c_2\\
&x_2=\int\psi_2(\tau)d\tau=-c_1\frac{t^2}{2}+c_2t+c_3\\
&x_1=\int x_2(\tau)d\tau=-c_1\frac{t^3}{6}+c_2\frac{t^2}{2}+c_3\,t+c_4
\end{aligned}$$

The optimal control law is thus finally:

$$u^*(t)=-\frac{c_1}{2K}t^2+\frac{c_2-c_1\tau_m}{K}t+\frac{c_3+c_2\tau_m}{K}\,,$$

where the constants c_1, c_2 and c_3 will be determined by the conditions imposed to the final state. The result is that the best control imposes to the scissors a time behavior which is a parabolic function, obtained by a control law itself parabolic.

Exercise 3.2 Optimal Control of the Double Integrator

Consider the following system described in state space by:

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = u \end{cases}$$

a) Calculate the optimal control law in stationary state which minimizes the cost functional

$$J = \frac{1}{2}\int_0^{\infty} (\rho^2 x_1^2 + u^2)\, dt, \text{ with } \rho > 0\,.$$

b) Determine the corresponding poles of the closed-loop system and comment its stability as a function of ρ.

Solution:

(a) The state equation of this system is of the form

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}u\,, \quad \mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

and the cost functional is the quadratic criterion

$$J = \frac{1}{2}\int_0^{\infty} (\mathbf{x}^{\mathrm{T}}\mathbf{Q}\,\mathbf{x} + \mathbf{u}^{\mathrm{T}}\mathbf{R}\,\mathbf{u})\, dt\,, \quad \mathbf{Q} = \begin{pmatrix} \rho^2 & 0 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{R} = 1\,.$$

Let us write the *algebraic* Riccati equation (3.54), which corresponds to the stationary case. We have, successively

$$\mathbf{B}^{\mathrm{T}}\mathbf{P} = \begin{pmatrix} 0 & 1 \end{pmatrix}\begin{pmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{pmatrix} = \begin{pmatrix} p_{12} & p_{22} \end{pmatrix}, \quad \mathbf{R}^{-1} = 1,$$

$$\mathbf{P}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P} = (\mathbf{B}^{\mathrm{T}}\mathbf{P})^{\mathrm{T}}(\mathbf{B}^{\mathrm{T}}\mathbf{P}) = \begin{pmatrix} p_{12} \\ p_{22} \end{pmatrix}\begin{pmatrix} p_{12} & p_{22} \end{pmatrix} = \begin{pmatrix} p_{12}^2 & p_{12}p_{22} \\ p_{12}p_{22} & p_{22}^2 \end{pmatrix},$$

$$\begin{pmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{pmatrix} - \begin{pmatrix} p_{12}^2 & p_{12}p_{22} \\ p_{12}p_{22} & p_{22}^2 \end{pmatrix} + \begin{pmatrix} \rho^2 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

This matrix equation is equivalent to the following system of three nonlinear equations:

$$\begin{cases} -p_{12}^2 + \rho^2 = 0 \\ p_{11} - p_{12}p_{22} = 0 \\ 2p_{12} - p_{22}^2 = 0 \end{cases}$$

Due to its quadratic nature, this system has more than one solution. It is easy to verify that the positive definite solution is given by:

$$\mathbf{P} = \begin{pmatrix} \rho\sqrt{2\rho} & \rho \\ \rho & \sqrt{2\rho} \end{pmatrix},$$

since the determinants associated with the upper left submatrices of **P**, $p_{11} = \rho\sqrt{2\rho}$ and $\begin{vmatrix} \rho\sqrt{2\rho} & \rho \\ \rho & \sqrt{2\rho} \end{vmatrix} = 2\rho^2 - \rho^2 = \rho^2$, are all strictly positive (see Appendix B, Sect. B.6).

The resulting control law is then:

$$u = -\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}\,\mathbf{x} = -\begin{pmatrix} 0 & 1 \end{pmatrix}\begin{pmatrix} \rho\sqrt{2\rho} & \rho \\ \rho & \sqrt{2\rho} \end{pmatrix} = -\begin{pmatrix} \rho & \sqrt{2\rho} \end{pmatrix}\mathbf{x} = -\boldsymbol{\ell}^{\mathrm{T}}\mathbf{x}.$$

(b) The closed-loop system matrix is:

$$\mathbf{A}_{CL} = \mathbf{A} - \mathbf{b}\boldsymbol{\ell}^{\mathrm{T}} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \end{pmatrix}\begin{pmatrix} \rho & \sqrt{2\rho} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -\rho & -\sqrt{2\rho} \end{pmatrix},$$

and the closed-loop characteristic polynomial is given by:

$$\det(s\mathbf{I} - \mathbf{A}_{CL}) = \begin{vmatrix} s & -1 \\ \rho & s + \sqrt{2\rho} \end{vmatrix} = s(s + \sqrt{2\rho}) + \rho = s^2 + \sqrt{2\rho}\, s + \rho\,.$$

The closed-loop poles are thus $s = \sqrt{\rho/2}\,(-1 \pm j)$, which characterizes a second order system with an undamped frequency $\omega_n = \sqrt{\rho}$ and a damping coefficient $\zeta = 1/\sqrt{2}$. Its stability is therefore guaranteed for any value of ρ, and its phase margin is superior to 60° (Sect. 3.7.1.2).

Exercise 3.3 Optimal Control of the Three-tank System

We consider here again the three-tank system introduced in Exercise 1.2. Design here an optimal control which brings the plant back to equilibrium after a disturbance of its initial state given by $x_0 = \begin{pmatrix}-0.8 & -0.5 & -0.2\end{pmatrix}^{\mathrm{T}}$, value already used in that exercise, in order to compare the results. Take for the two weighting matrices of the quadratic criterion the identity matrix in a first trial, and then another matrix **R** aiming at reducing the initial control.

Solution:

Run the program *MMCE.m* for this plant and choose the synthesis of a state feedback by the quadratic criterion (LQC) algorithm with state weighting.

With $\mathbf{Q} = \mathbf{I}_3$, $\mathbf{R} = \mathbf{I}_2$, the following control law is obtained:

$$\mathbf{L} = \begin{pmatrix} 0.776 & 0.331 & 0.106 \\ 0.106 & 0.283 & 0.620 \end{pmatrix}; \quad \mathbf{M} = \begin{pmatrix} 1.159 & 0.054 \\ 0.030 & 1.230 \end{pmatrix}.$$

The responses of the measured state variables $y_1 = x_1$ and $y_2 = x_3$, and the corresponding control signals, are reproduced in the left column of Fig. 3.5. Since one of the control signals of this trial comes close to the maximum value of the actuators (1 $\mathrm{m}^3 / \mathrm{mn}$), make a new trial with the choice $\mathbf{R} = 10 \times \mathbf{I}_2$, which yields

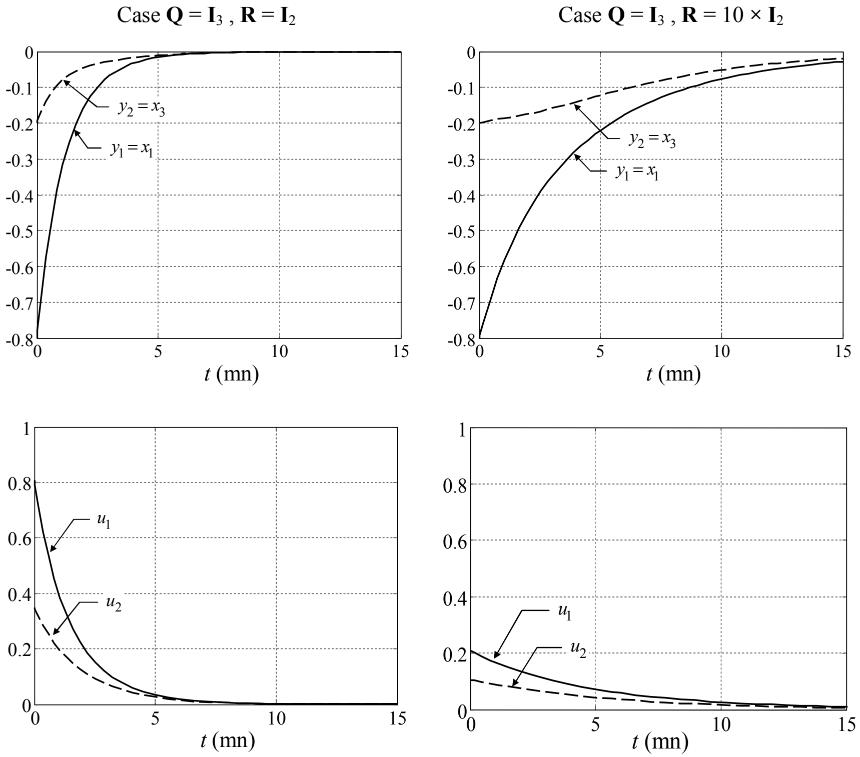
$$\mathbf{L} = \begin{pmatrix} 1.181 & 0.107 & 0.056 \\ 0.056 & 0.076 & 0.110 \end{pmatrix}; \quad \mathbf{M} = \begin{pmatrix} 0.452 & -0.107 \\ -0.123 & 0.617 \end{pmatrix}.$$

The responses of the same state variables and the associated controls are reproduced in the right column of Fig. 3.5.

The comparison with the plots of Exercise 1.2 shows that the dynamic responses are almost the same, especially in the first of the two cases studied here. In the second studied case, the controls are about four times weaker, but the response has become slower.

The advantage of the LQC control, as compared with the eigenvalue placement techniques, is the simplicity of the design. The drawback is that we have no control anymore over the dynamical behavior of the closed-loop, contrarily to the pole placement or modal control methods. This behavior will have to be adjusted in the case of quadratic control, by repetitive trials and simulations.

In Exercise 6.1 of Chap. 6, another design method will be described, which allows an LQC synthesis with regional pole constraint.


Case $\mathbf{Q} = \mathbf{I}_3$, $\mathbf{R} = \mathbf{I}_2$

Case $\mathbf{Q} = \mathbf{I}_3$, $\mathbf{R} = 10 \times \mathbf{I}_2$

Fig. 3.5 Free response of the state variables x_1 and x_3 (top) and associated controls (bottom).

Exercise 3.4 Optimal Control of the Inverted Pendulum

This exercise deals again with the inverted pendulum *LIP 100 (Amira)*, but this time we will design an optimal control for it.

a) Design an LQC optimal control, by choosing for the cost functional $J = \dfrac{1}{2}\displaystyle\int_0^{\infty} (\mathbf{x}^{\mathrm{T}} \mathbf{Q}\, \mathbf{x} + R u^2)\, dt$ successively the weighting matrices $\mathbf{Q} = 100 \times \mathbf{I}_4$, $R = 1$, and then $\mathbf{Q} = \mathbf{I}_4$, $R = 100$.

b) Verify by simulation that these two control laws stabilize the pendulum and compare the control signal in the two cases.

Solution:

(a) The LQC controller and gain compensation matrix are, respectively:
Case (1): $\mathbf{Q} = 100 \times \mathbf{I}_4$; $R = 1$:

$$\ell^{\mathrm{T}} = \begin{pmatrix} -10 & 34.28 & 32.37 & 12.40 \end{pmatrix};\ M = -10\,.$$

Case (2): $\mathbf{Q} = \mathbf{I}_4$; $R = 100$:

$$\ell^{\mathrm{T}} = \begin{pmatrix} -0.1 & 0.8686 & 0.8370 & 0.2119 \end{pmatrix};\ M = -0.1\,.$$

The step responses for a 1 volt position reference step applied at $t = 1$ s are given in Fig. 3.6 in the two cases.

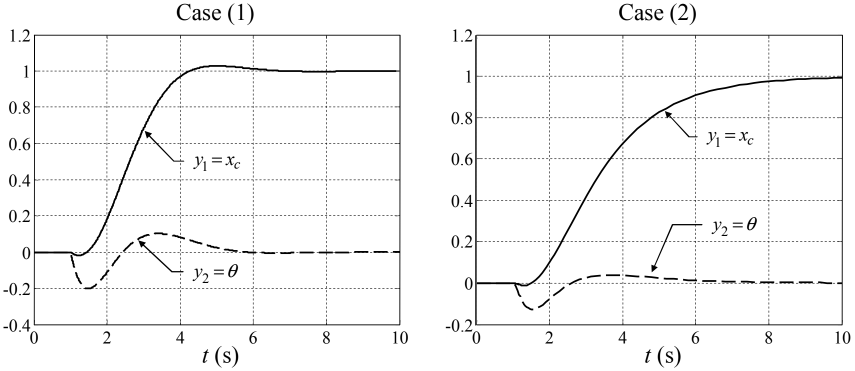


Fig. 3.6 Closed-loop step responses for the two LQC controllers.

(b) The control reaches -10 volts in case (1) and -0.1 volt in case (2). The choice of the weighting matrices favors significantly the energy savings in this second case, to the detriment of response speed.

Exercise 3.5 Optimal Acceleration of a Shopping Cart

Calculate what should be the optimal variation with time of the force $u(t)$ to be applied to a shopping cart, in the following conditions and hypotheses:

- the cart mass is taken as mass unity;
- the cart moves along a straight line, considered as the Ox axis, and is submitted to a viscous friction force $-k\dot{x}(t)$, where $x(t)$ is the cart position at time t ; the cart is initially at stop, at the origin of the Ox axis.

- the objective is to maximize the distance covered by the cart in a specified time t_f, while minimizing the total effort.

Solution:

This one-dimensional problem is described by the following motion equation:

$$\ddot{x}(t)+k\dot{x}(t)=u(t)\,,\quad x(0)=0\,,\ \dot{x}(0)=0\,.$$

The cost functional taking into account the optimization requirements is:

$$J=-x(t_f)+\int_0^{t_f}\frac{1}{2}u^2(t)\,dt\,.$$

A *minimum* value of J will thus correspond to a maximum value of $x(t_f)$ and to a minimum of effort spent all along the trajectory, as much for accelerating as for braking the cart.

With the state variables $x_1=x$ and $x_2=\dot{x}$, the system state representation is:

$$\dot{\mathbf{x}}=\begin{pmatrix}0 & 1\\ 0 & -k\end{pmatrix}\mathbf{x}+\begin{pmatrix}0\\1\end{pmatrix}u\,,\quad x_1(0)=x_2(0)=0\,, \tag{3.102}$$

and the cost functional becomes:

$$J=-x_1(t_f)+\int_0^{t_f}\frac{1}{2}u^2(t)\,dt\,.$$

This is a free final state, fixed final time problem, the basic problem of optimal control. With the notations of (3.2),

$$\begin{cases}h\left[x(t_f)\right]=-x_1(t_f)\\ f_0(\mathbf{x},u)=\dfrac{1}{2}u^2\end{cases}$$

By rewriting (3.102) in the form $\dot{\mathbf{x}}=\mathbf{f}(\mathbf{x},u)$, the components of $\mathbf{f}$ are identified:

$$\mathbf{f}=\begin{pmatrix}x_2\\ -kx_2+u\end{pmatrix}.$$

We calculate then the Hamiltonian, according to (3.4):

$$\boldsymbol{\psi}^{\mathrm{T}} \cdot \mathbf{f} = \begin{pmatrix} \psi_1 & \psi_2 \end{pmatrix} \begin{pmatrix} x_2 \\ -k x_2 + u \end{pmatrix} = \psi_1 x_2 - k\,\psi_2 x_2 + \psi_2 u\ ,$$

$$H = -\frac{1}{2} u^2 + \psi_1 x_2 + \psi_2 u - k\,\psi_2 x_2\ .$$

The optimal control law will be given by

$$\frac{\partial H}{\partial u} = 0 \quad \Rightarrow \quad -u + \psi_2 = 0 \quad \Rightarrow \quad u^*(t) = \psi_2(t)\ .$$

From the Hamiltonian expression the adjoint state equation is derived by (3.13):

$$\begin{cases} \dot{\psi}_1(t) = -\dfrac{\partial H}{\partial x_1} = 0 \\ \dot{\psi}_2(t) = -\dfrac{\partial H}{\partial x_2} = -\psi_1 + k\,\psi_2 \end{cases}$$

With the transversality condition (3.14),

$$\psi(t_f) = -\left.\frac{dh}{d\mathbf{x}}\right|_{t=t_f} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \psi_1(t_f) \\ \psi_2(t_f) \end{pmatrix},$$

the solution for the first component of the costate vector is $\psi_1(t) = \text{constant} = 1$, which yields for the second one

$$\dot{\psi}_2(t) - k\psi_2(t) = -\psi_1(t) = -1\ .$$

Since the expected solution has the form $\psi_2(t) = Ae^{kt} + C$ and the final condition is $\psi_2(t_f) = 0$, we obtain

$$\psi_2(t) = \frac{1}{k}\left[1 - e^{-k(t_f - t)}\right].$$

Hence the optimal control law:

$$u^*(t) = \frac{1}{k}\left[1 - e^{-k(t_f - t)}\right].$$

This law has a falling exponential shape, as illustrated in Fig. 3.7, which is plotted in arbitrary units and has been obtained for $t_f = 10$ and $k = 0.4$.

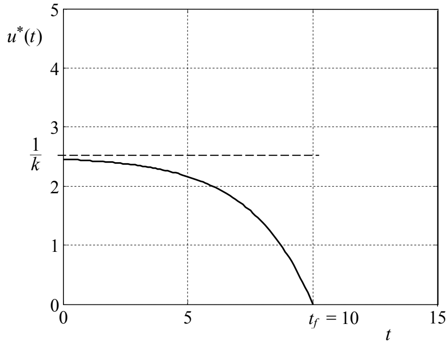


Fig. 3.7 Shape of the optimal control law to move the shopping cart.

Exercise 3.6 Discrete-time Optimal Control of an Air-Flow Heater

The plant used here is the "Process Trainer PT326" setup of the company Feedback™. It is an example of noisy system, and will be reused on grounds of that in the exercises of Chap. 4.

This process operates as a hair dryer: air is driven by a centrifugal blower through a tube, at the input of which it crosses a heater grid. The air temperature y, which represents the signal to be controlled, is measured by means of a thermistor inserted into the air stream at several points along the tube, here at its output. The control signal u is the voltage applied to the heater grid.

This plant is available in the *MMCE.m* program under the name process trainer PT326. Its continuous-time transfer function is (in rounded values):

$$G(s) = \frac{e^{-0.2s}}{(1+0.25s)^2}.$$

By sampling with zero-order hold at $T_s = 0.2$ second, the following discrete model is obtained, by the statement c2dm applied to the observability canonical model[2]:

[2] The state variables numbering of MATLAB® in the canonical forms is the inverse of that used in this book. This explains, among other, that $y_k = x_{1,k}$ and not $x_{2,k}$.

$$\mathbf{x}_{k+1} = \begin{pmatrix} 0.0899 & 0.0899 \\ -1.4379 & 0.8088 \end{pmatrix} \mathbf{x}_k + \begin{pmatrix} 0.1912 \\ 2.9675 \end{pmatrix} u_k$$

$$y_k = \begin{pmatrix} 1 & 0 \end{pmatrix} \mathbf{x}_k$$

In this exercise, the plant is to be controlled by optimal control.

a) Calculate the discrete-time optimal control law which minimizes the following quadratic criterion: $J = \sum_{k=0}^{\infty} (\mathbf{x}_k^{\mathrm{T}} \mathbf{Q}\, \mathbf{x}_k + R u_k^2)$, by taking $\mathbf{Q} = \mathbf{I}_2$ and $R = 100$.

b) Repeat this design with other ratios between the two weighting matrices, such as $\mathbf{Q} = 100 \times \mathbf{I}_2$; $R = 1$. Observe by simulation the effect of these choices on the step responses of the output and on the control signal.

c) Verify the robustness of this control law by using the approach in Sect. 3.7.2 and by plotting the Nyquist locus of the compensated open loop, for $\mathbf{Q} = \mathbf{I}_2$; $R = 10$.

Solution:

(a) Run the *MMCE.m* program by choosing successively the plant to study #3, then discrete model, synthesis of state feedback, without integral action, quadratic criterion algorithm (LQC) with state weighting, and introduce the values of the exercise statement.

The following control law is obtained:

$$\boldsymbol{\ell}^{\mathrm{T}} = \begin{pmatrix} -0.0691 & 0.0335 \end{pmatrix}; \; M = 1.1993 .$$

The step response of the measured output $y = x_1$ and the control signal u are recorded by simulation and plotted on the left side of Fig. 3.8.

(b) The control law is now:

$$\boldsymbol{\ell}^{\mathrm{T}} = \begin{pmatrix} -0.4796 & 0.2732 \end{pmatrix}; \; M = 2.7057 ,$$

and the output step response and control signal are reproduced on the right side of Fig. 3.8. It is seen that the transient response is faster (about 0.5 s rise time, versus 1 s in the first case), but at the cost of a much more energetic control.

This is in agreement with the chosen weightings: in this second case, indeed, we have required small trajectory deviations (large coefficients in **Q**), while we did not require control energy minimization (small R). The first case is the exact opposite of this one.

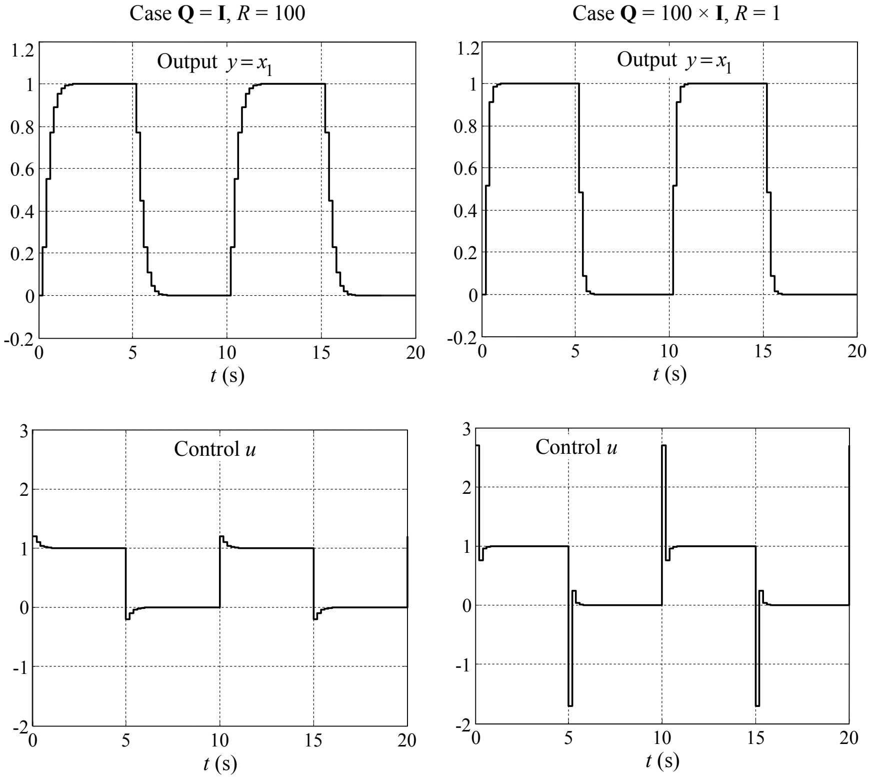


Fig. 3.8 Step responses and corresponding control signals for questions (a) and (b).

(c) The closed-loop poles are $0.4062 \pm j0.1049$ in the first case, and 0.1785 and 0.0013 in the second one.

The closed-loop system is thus stable. To check its robustness, plot with the script file *LQ_LQG_robustness_discrete.m*, contained in the downloadable *mmce.zip* archive, the Nyquist diagram of the compensated open loop and the disk which it should avoid, according to (3.92) in Sect. 3.7.2.2. As shown in Fig. 3.9, this is indeed what happens.

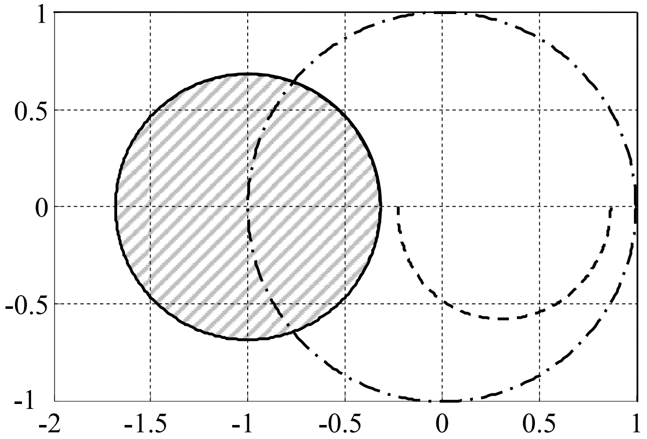
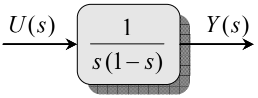


Fig. 3.9 Compensated open-loop Nyquist diagram (dashed line), forbidden disk (hatched), and unit circle (dash-dotted line).

Exercise 3.7 Optimal Control by Solving the ARE.

Consider the plant



a) Determine the state space representation of this plant in diagonal form.

b) Calculate the state-feedback optimal control law, which minimizes the criterion $J = \dfrac{1}{2}\displaystyle\int_0^\infty (y^2 + u^2)dt$.

c) Study the robustness of this control law by following the approach of Sect. 3.7.1.

Solution:

(a) By partial fraction expansion, the above transfer function can be decomposed as

$$\frac{Y(s)}{U(s)} = \frac{1}{s(1-s)} = \frac{1}{s} - \frac{1}{s-1},$$

and the following diagonal state representation is obtained readily:

$$\begin{cases} \dot{\mathbf{x}} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \mathbf{x} + \begin{pmatrix} 1 \\ 1 \end{pmatrix} u \\ y = \begin{pmatrix} 1 & -1 \end{pmatrix} \mathbf{x} \end{cases}$$

(b) According to (3.28) we have

$$f_0 = \frac{1}{2}(y^2 + u^2) = \frac{1}{2}(\mathbf{x}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}\mathbf{C}\mathbf{x} + uRu) = \frac{1}{2}(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + uRu),$$

where

$$\mathbf{Q} = \mathbf{C}^{\mathrm{T}}\mathbf{C} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \begin{pmatrix} 1 & -1 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}; \; R = 1 .$$

The algebraic Riccati equation (3.54) is thus, with **P** (symmetric) $= \begin{pmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{pmatrix}$:

$$\begin{pmatrix} 0 & p_{12} \\ 0 & p_{22} \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ p_{12} & p_{22} \end{pmatrix} - \begin{pmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}.$$

It splits into the following three equations:

$$\begin{cases} -p_{11}(p_{11} + p_{12}) - p_{12}(p_{11} + p_{12}) = -1 \\ p_{12} - (p_{11} + p_{12})(p_{12} + p_{22}) = 1 \\ 2p_{22} - (p_{12} + p_{22})(p_{12} + p_{22}) = -1 \end{cases}$$

By substitution and elimination, the positive definite solution of this system of quadratic equations is determined:

$$\mathbf{P} = \begin{pmatrix} \sqrt{3} & -1-\sqrt{3} \\ -1-\sqrt{3} & 3+2\sqrt{3} \end{pmatrix}.$$

The optimal controller in stationary state is thus:

$$\mathbf{L} = \boldsymbol{\ell}^{\mathrm{T}} = R^{-1}\mathbf{b}^{\mathrm{T}}\mathbf{P} = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{3} & -1-\sqrt{3} \\ -1-\sqrt{3} & 3+2\sqrt{3} \end{pmatrix} = \begin{pmatrix} -1 & 2+\sqrt{3} \end{pmatrix}.$$

N.B.: this solution can also be obtained numerically by the MATLAB® statement lqry(A,B,C,0,Q,R), which uses directly a quadratic form of the output, $\mathbf{y}^T\mathbf{Q}\mathbf{y}$.

(c) The Nyquist locus of the compensated open loop, plotted by means of the script file *LQ_robustness_continuous.m* available in the downloadable *mmce.zip* archive, is shown in Fig. 3.10.

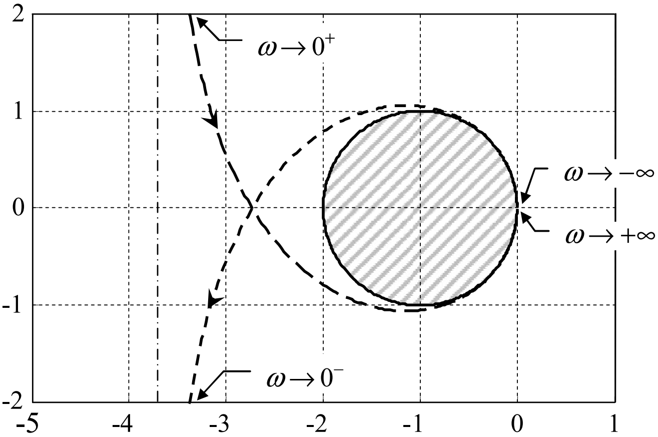


Fig. 3.10 Nyquist diagram of the compensated open loop (continuous case).

As expected, the Nyquist locus does not enter the unit radius disk centered at -1. This guarantees the closed loop an infinite gain margin and a phase margin of at least 60° (see Sect. 3.7.1.2).

4 Noisy Systems – Optimal Linear Filtering

This chapter deals with noise signals, also called *random* or *stochastic processes*, which affect the plant itself or the available measurements. As mentioned earlier, the state estimator will be here a *filter*. The optimal linear filter, if optimization is understood as *minimization of the estimation error variance*, is the well known Kalman filter. This chapter is organized as follows: after an introduction to the estimation theory, the description of noisy systems in state space will be presented. The obtained equations will then be used to establish the Kalman filter, first for the discrete case, which is easier to apprehend in an initial approach than the continuous case, mainly for notation reasons, and then also for the continuous case. The choice of the noise covariance matrices will be discussed after that, and the use of a Kalman filter as an observer in the control loop will be presented. Though this book is entirely devoted to linear systems, a short digression will occur about an extended version of the Kalman filter, which applies to nonlinear systems.

All over this chapter the duality with the optimal control approach of Chap. 3 will be emphasized and lead to the *Duality Principle*, which reaches far beyond the questions discussed here.

The basic notions of Probability Theory, for scalar or vector signals, and the notations used in the present chapter are summarized in Appendix C. More details can be found in textbooks about this topic such as [Kay06] or [ShBr88].

4.1 General Theory of Estimation

The problem of estimating the state of a noisy system, thus of estimating a random vector, will be cast first in the context of the estimation theory.

4.1.1 Introduction to the Estimation Theory

It is useful here to recall a few basic notions, in order to highlight the differences between the deterministic and the stochastic situations.

E. Ostertag, *Mono- and Multivariable Control and Estimation,*
DOI 10.1007/978-3-642-13734-1_4, © Springer-Verlag Berlin Heidelberg 2011

4.1.1.1 Unknowns, Measurement, Estimation

Two categories of variables will be considered in the sequel, the word *variable* being used in the large sense which encompasses as well scalar as vector variables:
x: the unknown variables to estimate
y: the measured variables, whose value depends in some way on that of **x**.

- **x** can be:
  - either the vector of parameters of a system to be identified: $\mathbf{x} = \boldsymbol{\theta}$; one speaks then of *system identification*; e.g. if a SISO discrete-time system is to be identified in terms of a *z* transfer function,

$$G(z) = \frac{b_0 + b_1 z^{-1} + \ \dots \ + b_n z^{-n}}{1 + a_1 z^{-1} + \ \dots \ + a_n z^{-n}},$$

    the vector of parameters will be: $\boldsymbol{\theta} = \begin{pmatrix} a_1 & a_2 & \cdots & a_n & b_0 & b & \cdots & b_n \end{pmatrix}^{\mathrm{T}}$;
  - or the set of successive states of a system, e.g. in case of a discrete-time system, the extended vector $\mathbf{x}^{\mathrm{T}} = \begin{pmatrix} \mathbf{x}_1^{\mathrm{T}} & \mathbf{x}_2^{\mathrm{T}} & \cdots & \mathbf{x}_k^{\mathrm{T}} & \cdots \end{pmatrix}$, or simply the system state vector $\mathbf{x} = \mathbf{x}_k$ at time $t = kT$. In this case one will speak of *state estimation*, or *reconstruction*.
- **y** will be the set of made measurements, e.g. in the case of a discrete-time system the input-output measurement

$$\mathbf{y}^{\mathrm{T}} = \begin{pmatrix} \mathbf{y}_1^{\mathrm{T}} & \mathbf{y}_2^{\mathrm{T}} & \cdots & \mathbf{y}_k^{\mathrm{T}} & \cdots & \mathbf{u}_1^{\mathrm{T}} & \mathbf{u}_2^{\mathrm{T}} & \cdots & \mathbf{u}_k^{\mathrm{T}} & \cdots \end{pmatrix}.$$

  In general, the measurements **y** depend on the value of **x**. This dependency is
  - *deterministic*, if **y** is completely determined by **x;**
  - *stochastic*, if **y** depends not only on **x**, but also on chance due to noise or unknown disturbances.

In summary, the situation can be sketched as follows:

$$\text{unknowns } \mathbf{x} \xrightarrow[]{\substack{\text{deterministic or stochastic} \\ \text{dependence}}} \text{measurements } \mathbf{y}.$$

To estimate **x** as a function of **y** consists in manipulating these measurements to obtain a quantity $\widehat{\mathbf{x}}$ called *estimation,* or *estimate,* of **x**:

$$\text{measurements } \mathbf{y} \xrightarrow{\text{deterministic dependence}} \text{estimate } \widehat{\mathbf{x}}$$

The considered manipulation is deterministic in the sense that, to a given vector **y**, corresponds a deterministic value of $\widehat{\mathbf{x}}$. An estimate is thus a *function* of **y:**

$$\widehat{\mathbf{x}} = \mathbf{g}(\mathbf{y}) .$$

There are many identification or estimation methods, thus many estimation functions $\mathbf{g}(\mathbf{y})$, which will yield different estimations $\widehat{\mathbf{x}}$, more or less satisfactory, the degree of satisfaction depending of course on the closeness of $\widehat{\mathbf{x}}$ to **x**.

To appreciate whether a given function **g** is satisfactory, it will be necessary to adopt a statistical viewpoint, and to consider that part or all of the variables (**x**, **y**) involved in the problem are *realizations* of *random variables*.

4.1.1.2 Variables to Estimate and Measurements, as Realizations of Random Variables

Le us consider indeed the experiment which consists in recording the inputs and the outputs of a given plant type, for its identification. One such record is $\mathbf{y}^{\mathrm{T}} = \begin{pmatrix} \mathbf{u}_0^{\mathrm{T}} & \mathbf{u}_1^{\mathrm{T}} & \cdots & \mathbf{u}_n^{\mathrm{T}} & \mathbf{y}_0^{\mathrm{T}} & \mathbf{y}_1^{\mathrm{T}} & \cdots & \mathbf{y}_n^{\mathrm{T}} \end{pmatrix}$. By repeating the experiment, a different vector **y** will be obtained, for one reason because of measurement noise, but also because the input sequence may well be random from one test to the next one, as is the case in system identification by pseudo-random binary sequences:

- one will therefore say that **y** is a realization of a random vector **Y**;
- consequently, given some estimation function **g**, $\widehat{\mathbf{x}} = \mathbf{g}(\mathbf{y})$ will be a realization of a random vector $\dot{\mathbf{X}} = \mathbf{g}(\mathbf{Y})$.

Remark 4.1. As is common practice in Probability Theory, lower case letters denote the realizations, and capital letters the random quantities (see also Notational Conventions, at the beginning of this book).

The random variable $\dot{\mathbf{X}}$ is called *estimator*. The realization $\widehat{\mathbf{x}}$ is called *estimation* or *estimate*.

As far as the true vector **x** (parameters, unknown states) is concerned, two situations can be envisioned:

- **1st case**: **x** is a realization of a *random variable* **X**. The corresponding theory is called *general estimation theory*.

  This is the case, e.g., when one wants to estimate the parameters of a class of systems or to estimate the state of a given system, which is corrupted with process noise: if the experiment were repeated, **x** would present each time a different value.

This is the situation which will prevail for the remainder of this chapter, and it is thus the general estimation theory which will be developed here.

- **2nd case**: $\mathbf{x}$ is an unknown *deterministic* quantity. The relevant theory is called *classical estimation theory.*

  This is the case of the identification of a given system, where all successive measurement data $\mathbf{y}$ are recorded on the same system with the same state vector $\mathbf{x}$, or of the estimation of the state of a system immune to process noise.

4.1.1.3 Estimation Error

In the two previous cases, the estimation error will be defined as the difference between real value and its estimate: $\tilde{\mathbf{x}} = \mathbf{x} - \hat{\mathbf{x}}$. This error, unknown if $\mathbf{x}$ is unknown, is always a *realization of a random variable* $\tilde{\mathbf{X}}$. Indeed, according to the situation envisioned, the following will hold:

$$\tilde{\mathbf{X}} = \mathbf{X} - \dot{\mathbf{X}} \quad \text{(general theory)},$$

$$\tilde{\mathbf{X}} = \mathbf{x} - \dot{\mathbf{X}} \quad \text{(classical theory)}.$$

To evaluate the quality of a given estimation function $\mathbf{g}$, $\mathrm{E}\left\{\tilde{\mathbf{X}}\right\}$ is often used, i.e. the mathematical expectation of estimation errors which would be observed in a large number of experiments, growing at the limit towards infinity.

4.1.2 Linear, Unbiased, Minimum Error-variance Estimate

Consider two random vectors, $\mathbf{X}$ of length n and $\mathbf{Y}$ of length q.

4.1.2.1 Linear Estimation

One calls *linear estimation* of $\mathbf{X}$ from $\mathbf{Y}$ an estimation $\hat{\mathbf{x}} = \mathbf{g}(\mathbf{y})$ for which $\mathbf{g}$ is a linear function of $\mathbf{y}$:

$$\hat{\mathbf{x}} = \mathbf{K}\,\mathbf{y} + \boldsymbol{\ell} \; . \tag{4.1}$$

$\hat{\mathbf{x}}$ is then a realization of the *linear estimator*

$$\dot{\mathbf{X}} = \mathbf{K}\mathbf{Y} + \boldsymbol{\ell} \ . \tag{4.2}$$

4.1.2.2 Unbiased Linear Estimation

Let $\tilde{\mathbf{X}} = \mathbf{X} - \dot{\mathbf{X}}$ be the estimation error. An estimation is said *unbiased,* if

$$\mathrm{E}\left\{\tilde{\mathbf{X}}\right\} = 0 \ . \tag{4.3}$$

The linear estimate (4.1) will thus be unbiased if

$$\mathrm{E}\left\{\tilde{\mathbf{X}}\right\} = \mathrm{E}\left\{\mathbf{X} - \dot{\mathbf{X}}\right\} = \mathrm{E}\left\{\mathbf{X} - \mathbf{K}\mathbf{Y} - \boldsymbol{\ell}\right\} = \boldsymbol{\mu}_X - \mathbf{K}\boldsymbol{\mu}_Y - \boldsymbol{\ell} = 0 \,,$$

where $\boldsymbol{\mu}_X = \mathrm{E}\left\{\mathbf{X}\right\}$, $\boldsymbol{\mu}_Y = \mathrm{E}\left\{\mathbf{Y}\right\}$, i.e. if $\boldsymbol{\ell} = \boldsymbol{\mu}_X - \mathbf{K}\boldsymbol{\mu}_Y$. The *unbiased linear estimate* of **X** from **Y** is thus

$$\hat{\mathbf{x}} = \boldsymbol{\mu}_X + \mathbf{K}(\mathbf{y} - \boldsymbol{\mu}_Y) \ , \tag{4.4}$$

corresponding, of course, to the estimator

$$\dot{\mathbf{X}} = \boldsymbol{\mu}_X + \mathbf{K}(\mathbf{Y} - \boldsymbol{\mu}_Y) \ . \tag{4.5}$$

4.1.2.3 Optimal (Unbiased) Linear Estimate

Let us look now, in the class of unbiased linear estimations, for an estimation which minimizes the estimation error variance.

Since $\mathrm{E}\left\{\tilde{\mathbf{X}}\right\} = 0$, the estimation error variance is by definition the covariance matrix

$$\boldsymbol{\Sigma} = \mathrm{Var}\left\{\tilde{\mathbf{X}}\right\} = \mathrm{E}\left\{\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathrm{T}}\right\} = \mathrm{E}\left\{(\mathbf{X} - \dot{\mathbf{X}})(\mathbf{X} - \dot{\mathbf{X}})^{\mathrm{T}}\right\} .$$

According to (4.5),

$$\begin{aligned}
\boldsymbol{\Sigma} &= \mathrm{E}\left\{\left[(\mathbf{X} - \boldsymbol{\mu}_X) - \mathbf{K}(\mathbf{Y} - \boldsymbol{\mu}_Y)\right]\left[(\mathbf{X} - \boldsymbol{\mu}_X) - \mathbf{K}(\mathbf{Y} - \boldsymbol{\mu}_Y)\right]^{\mathrm{T}}\right\} \\
&= \mathrm{E}\left\{(\mathbf{X} - \boldsymbol{\mu}_X)(\mathbf{X} - \boldsymbol{\mu}_X)^{\mathrm{T}}\right\} - \mathrm{E}\left\{(\mathbf{X} - \boldsymbol{\mu}_X)(\mathbf{Y} - \boldsymbol{\mu}_Y)^{\mathrm{T}}\right\}\mathbf{K}^{\mathrm{T}} \\
&\qquad - \mathbf{K}\,\mathrm{E}\left\{(\mathbf{Y} - \boldsymbol{\mu}_Y)(\mathbf{X} - \boldsymbol{\mu}_X)^{\mathrm{T}}\right\} + \mathbf{K}\,\mathrm{E}\left\{(\mathbf{Y} - \boldsymbol{\mu}_Y)(\mathbf{Y} - \boldsymbol{\mu}_Y)^{\mathrm{T}}\right\}\mathbf{K}^{\mathrm{T}} ,
\end{aligned}$$

thus, using the notations of (C.9) and (C.11) in Appendix C,

$$\mathbf{\Sigma} = \mathbf{\Sigma}_{XX} - \mathbf{\Sigma}_{XY}\mathbf{K}^{\mathrm{T}} - \mathbf{K}\mathbf{\Sigma}_{YX} + \mathbf{K}\mathbf{\Sigma}_{YY}\mathbf{K}^{\mathrm{T}}\,. \tag{4.6}$$

The problem consists in finding the $(n \times q)$ matrix $\mathbf{K}$ which minimizes this error variance $\mathbf{\Sigma}$. As a matter of fact, since we do not know how to minimize a matrix, and owing to the particular structure of $\mathbf{\Sigma}$ (see (C.9)), we will look for the matrix $\mathbf{K}$ which minimizes the sum of the estimation error variances of all the components of $\mathbf{X}$, i.e. the following criterion:

$$J = \sum_{i=1}^{n} \mathrm{Var}\left\{\tilde{X}_i\right\} = \sum_{i=1}^{n} \mathrm{E}\left\{\tilde{X}_i^2\right\} = \mathrm{E}\left\{\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}}\right\}. \tag{4.7}$$

By application of (B.2) of Appendix B,

$$\mathrm{E}\left\{\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}}\right\} = \mathrm{E}\left\{\mathrm{tr}\left[\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathrm{T}}\right]\right\} = \mathrm{tr}\left[\mathrm{E}\left\{\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathrm{T}}\right\}\right],$$

so that, finally:

$$J = \mathrm{tr}(\mathbf{\Sigma})\,.$$

According to (4.6),

$$\mathrm{tr}(\mathbf{\Sigma}) = \mathrm{tr}(\mathbf{\Sigma}_{XX}) - \mathrm{tr}(\mathbf{\Sigma}_{XY}\mathbf{K}^{\mathrm{T}}) - \mathrm{tr}(\mathbf{K}\mathbf{\Sigma}_{YX}) + \mathrm{tr}(\mathbf{K}\mathbf{\Sigma}_{YY}\mathbf{K}^{\mathrm{T}})\,.$$

Since, according to (C.12) of Appendix C, $\mathbf{\Sigma}_{YX}^{\mathrm{T}} = \mathbf{\Sigma}_{XY}$, the third term on the right side of this equation evaluates as

$$\mathrm{tr}(\mathbf{K}\mathbf{\Sigma}_{YX}) = \mathrm{tr}(\mathbf{K}\mathbf{\Sigma}_{YX})^{\mathrm{T}} = \mathrm{tr}(\mathbf{\Sigma}_{YX}^{\mathrm{T}}\mathbf{K}^{\mathrm{T}}) = \mathrm{tr}(\mathbf{\Sigma}_{XY}\mathbf{K}^{\mathrm{T}})\,.$$

Hence:

$$\mathrm{tr}(\mathbf{\Sigma}) = \mathrm{tr}(\mathbf{\Sigma}_{XX}) - 2\,\mathrm{tr}(\mathbf{\Sigma}_{XY}\mathbf{K}^{\mathrm{T}}) + \mathrm{tr}(\mathbf{K}\mathbf{\Sigma}_{YY}\mathbf{K}^{\mathrm{T}})\,.$$

By applying the differentiation relations of a scalar with respect to a matrix, (B.19) and (B.21) of Appendix B, we obtain

$$\frac{d}{d\mathbf{K}}\mathrm{tr}(\mathbf{\Sigma}) = -2\,\mathbf{\Sigma}_{XY} + 2\,\mathbf{K}\mathbf{\Sigma}_{YY}\,,$$

since $\mathbf{\Sigma}_{YY}$ is symmetric (see (C.10), Appendix C). The minimum of $\mathrm{tr}(\mathbf{\Sigma})$ is obtained by cancelling this derivative, i.e. for $\mathbf{K}$ solution of

$$\mathbf{K}\boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{XY} = 0 \,,$$

yielding

$$\mathbf{K} = \boldsymbol{\Sigma}_{XY}\,\boldsymbol{\Sigma}_{YY}^{-1} \ . \tag{4.8}$$

Substituting now this value in (4.6), we obtain for $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{XX} - \mathbf{K}\boldsymbol{\Sigma}_{YX} = \boldsymbol{\Sigma}_{XX} - \boldsymbol{\Sigma}_{XY}\,\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX} \ .$$

Let us gather the previous results together.

Theorem 4.1. *If* $\mathbf{X}$ *and* $\mathbf{Y}$ *are two random vectors, with variances and covariance* $\boldsymbol{\Sigma}_{XX}$*,* $\boldsymbol{\Sigma}_{YY}$ *and* $\boldsymbol{\Sigma}_{XY}$*, the unbiased, linear estimate of* $\mathbf{X}$ *from* $\mathbf{Y}$*, which is optimal in the sense of the minimum variance of* $\widetilde{\mathbf{X}} = \mathbf{X} - \widehat{\mathbf{X}}$*, is defined by*

$$\widehat{\mathbf{x}} = \boldsymbol{\mu}_X + \boldsymbol{\Sigma}_{XY}\,\boldsymbol{\Sigma}_{YY}^{-1}(\mathbf{y} - \boldsymbol{\mu}_Y) \ , \tag{4.9}$$

estimate for which

$$\boldsymbol{\Sigma} = \mathrm{Var}\left\{\widetilde{\mathbf{X}}\right\} = \boldsymbol{\Sigma}_{XX} - \boldsymbol{\Sigma}_{XY}\,\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX} \ . \tag{4.10}$$

Remark 4.2. The optimal linear *estimator* corresponding to (4.9) is given by

$$\dot{\mathbf{X}} = \boldsymbol{\mu}_X + \boldsymbol{\Sigma}_{XY}\,\boldsymbol{\Sigma}_{YY}^{-1}(\mathbf{Y} - \boldsymbol{\mu}_Y) \ . \tag{4.11}$$

Remark 4.3. By developing (4.7),

$$J = \sum_{i=1}^{n} \mathrm{E}\left\{\widetilde{X}_i^2\right\} = \mathrm{E}\left\{\sum_{i=1}^{n} \widetilde{X}_i^2\right\} = \mathrm{E}\left\{\sum_{i=1}^{n} (X_i - \widehat{X}_i)^2\right\},$$

it appears that the above minimization amounts to minimizing the sum of the squares of the differences between the various components of $\mathbf{X}$ and of $\dot{\mathbf{X}}$, in stochastic average (average norm of the error vector). This is why the linear optimal estimate obtained above is also said *optimal in the sense of the least squares*.

4.1.3 Orthogonality Principle

We saw that the minimum error-variance, linear, estimator, which is of the form (4.11), was unbiased, thus that $\mathrm{E}\left\{\widetilde{\mathbf{X}}\right\} = 0$. Let us prove now that, in addition,

$$\mathrm{Cov}\left\{\tilde{\mathbf{X}},\mathbf{Y}\right\}=0\,,$$

i.e. that the measurement $\mathbf{Y}$ and the estimation error $\tilde{\mathbf{X}}=\mathbf{X}-\dot{\mathbf{X}}$ are *uncorrelated.*

The following holds readily:

$$\begin{aligned}\mathrm{Cov}\left\{\tilde{\mathbf{X}},\mathbf{Y}\right\}&=\mathrm{E}\left\{\left(\tilde{\mathbf{X}}-\mathrm{E}\left\{\tilde{\mathbf{X}}\right\}\right)\left(\mathbf{Y}-\boldsymbol{\mu}_Y\right)^{\mathrm{T}}\right\}=\mathrm{E}\left\{(\mathbf{X}-\dot{\mathbf{X}})(\mathbf{Y}-\boldsymbol{\mu}_Y)^{\mathrm{T}}\right\}\\&=\mathrm{E}\left\{\left[(\mathbf{X}-\boldsymbol{\mu}_X)-\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}(\mathbf{Y}-\boldsymbol{\mu}_Y)\right]\left[\mathbf{Y}-\boldsymbol{\mu}_Y\right]^{\mathrm{T}}\right\}=\boldsymbol{\Sigma}_{XY}-\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YY}=0\,.\end{aligned}$$

On the other hand,

$$\mathrm{Cov}\left\{\tilde{\mathbf{X}},\mathbf{Y}\right\}=\mathrm{E}\left\{\tilde{\mathbf{X}}(\mathbf{Y}-\boldsymbol{\mu}_Y)^{\mathrm{T}}\right\}=\mathrm{E}\left\{\tilde{\mathbf{X}}\mathbf{Y}^{\mathrm{T}}\right\}-\mathrm{E}\left\{\tilde{\mathbf{X}}\right\}\boldsymbol{\mu}_Y^{\mathrm{T}}=\mathrm{E}\left\{\tilde{\mathbf{X}}\mathbf{Y}^{\mathrm{T}}\right\}.$$

The comparison of these two equations implies that $\mathrm{E}\left\{\tilde{\mathbf{X}}\mathbf{Y}^{\mathrm{T}}\right\}=0$, thus that $\tilde{\mathbf{X}}$ and $\mathbf{Y}$ are also *orthogonal.*

Reciprocally, let us prove that if a linear estimator $\dot{\mathbf{X}}=\mathbf{K}\mathbf{Y}+\boldsymbol{\ell}$ is:

1. unbiased,
2. such that $\mathrm{Cov}\left\{\tilde{\mathbf{X}},\mathbf{Y}\right\}=0$,

it is then an optimal linear estimator.

The unbiased property implies, as proved in Sect. 4.1.2, that it is of the form (4.5), namely

$$\dot{\mathbf{X}}=\boldsymbol{\mu}_X+\mathbf{K}(\mathbf{Y}-\boldsymbol{\mu}_Y)\,.$$

Furthermore, the fact that $\mathrm{Cov}\left\{\tilde{\mathbf{X}},\mathbf{Y}\right\}=0$ yields that

$$\begin{aligned}0=\mathrm{Cov}\left\{\tilde{\mathbf{X}},\mathbf{Y}\right\}&=\mathrm{E}\left\{(\mathbf{X}-\dot{\mathbf{X}})(\mathbf{Y}-\boldsymbol{\mu}_Y)^{\mathrm{T}}\right\}\\&=\mathrm{E}\left\{\left[(\mathbf{X}-\boldsymbol{\mu}_X)-\mathbf{K}(\mathbf{Y}-\boldsymbol{\mu}_Y)\right]\left[\mathbf{Y}-\boldsymbol{\mu}_Y\right]^{\mathrm{T}}\right\}=\boldsymbol{\Sigma}_{XY}-\mathbf{K}\,\boldsymbol{\Sigma}_{YY}\,.\end{aligned}$$

Thus: $\mathbf{K}=\boldsymbol{\Sigma}_{XY}\,\boldsymbol{\Sigma}_{YY}^{-1}$. The optimal value of $\mathbf{K}$ is regained here.

Remark 4.4. Here again it came to the same to assume $\tilde{\mathbf{X}}$ and $\mathbf{Y}$ orthogonal as to assume them uncorrelated. If indeed $\mathrm{E}\left\{\tilde{\mathbf{X}}\mathbf{Y}^{\mathrm{T}}\right\}=0$, which is the condition of orthogonality, the following is true:

$$\mathrm{Cov}\left\{\tilde{\mathbf{X}},\mathbf{Y}\right\}=\mathrm{E}\left\{\tilde{\mathbf{X}}(\mathbf{Y}-\boldsymbol{\mu}_Y)^{\mathrm{T}}\right\}=\mathrm{E}\left\{\tilde{\mathbf{X}}\mathbf{Y}^{\mathrm{T}}\right\}-\mathrm{E}\left\{\tilde{\mathbf{X}}\right\}\boldsymbol{\mu}_Y^{\mathrm{T}}=0\,,$$

since the estimator was assumed unbiased. $\tilde{\mathbf{X}}$ and **Y** are thus uncorrelated.

Conclusion: Principle of Orthogonality. *The uncorrelation of* $\tilde{\mathbf{X}}=\mathbf{X}-\dot{\mathbf{X}}$ *and of* $\mathbf{Y}$*, or their orthogonality, is thus a necessary and sufficient condition of the optimality of* $\dot{\mathbf{X}}$.

Physical meaning of the principle of orthogonality. If **Y** enables estimating **X**, it is because there exists a correlation between **X** and **Y**. If this were not the case, the consequence would be indeed that $\boldsymbol{\Sigma}_{XY}=0$, and the estimation would be reduced to $\dot{\mathbf{X}}=\boldsymbol{\mu}_X$. If now a correlation were remaining between $(\mathbf{X}-\dot{\mathbf{X}})$ and **Y**, we could thus estimate the difference $(\mathbf{X}-\dot{\mathbf{X}})$, and therefore improve the estimation $\dot{\mathbf{X}}$. This would mean that the estimate was not optimal. Thus, if no correlation subsists between $(\mathbf{X}-\dot{\mathbf{X}})$ and **Y**, it means that $\dot{\mathbf{X}}$ is optimal, and vice-versa.

4.2 Statistical Relations in State Space

4.2.1 Introduction

The state space representation of systems, considered up to now deterministic, will be extended in this section to the description of stochastic processes.

More precisely, it will be shown that every *Markovian* stochastic process, as defined hereafter, can be considered as the solution of a stochastic difference or differential equation, according to its nature (discrete-time or continuous-time), and will thus be modelizable by the state of a linear system fed by white noise.

4.2.2 Markovian Stochastic Processes

A stochastic process $\mathbf{X}(u)$ is Markovian if its evolution after an instant t depends only on the value of $\mathbf{x}(t)$ at that time, and remains thus independent of its value at times $t_i<t$.

Definition 4.1. A continuous or discrete stochastic process $\mathbf{X}(t)$ is said Markovian if, for an arbitrary sequence of times such as $t_k < \cdots < t_1 < t < u$, the conditional random variable

$$\mathbf{X}(u) \,\big|\, \mathbf{X}(t), \mathbf{X}(t_1), \ldots, \mathbf{X}(t_k)$$

has the same probability distribution as the conditional random variable

$$\mathbf{X}(u) \,\big|\, \mathbf{X}(t) \, .$$

Remark 4.5. A Markovian process is also called a memoryless process.

4.2.3 Discrete-time Markovian Process. Noisy Discrete Systems

4.2.3.1 Stochastic State. Stochastic Linear Difference Equation

Let us consider a discrete-time linear system, at the input of which a vectorial random signal $\mathbf{V}_k$ is applied instead of a deterministic signal.

The state of the discrete system becomes then a *stochastic state* $\mathbf{X}_k$ and its difference state equation becomes a *stochastic linear difference state equation*:

$$\mathbf{X}_{k+1} = \mathbf{\Phi}_k \mathbf{X}_k + \mathbf{E}_k \mathbf{V}_k \ . \tag{4.12}$$

Remark 4.6. This state equation is considered here in the general case of a linear system, not necessarily time-invariant. The matrix

$$\mathbf{\Phi}_k = \mathbf{\Phi}(t_{k+1}, t_k)$$

is the transition matrix of this system from step k to step $k+1$, and therefore depends here on time (see Sect. A.5, Appendix A). For the sake of generality, the same assumption will be made concerning $\mathbf{E}_k$. Such a system is represented in state space form in Fig. 4.1.

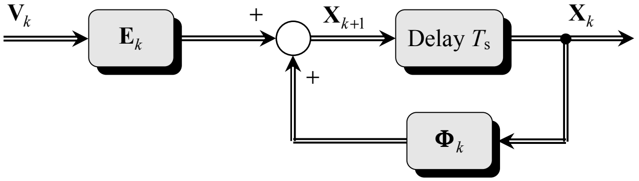


Fig. 4.1 Time-varying, linear, discrete-time system with random input and random state.

4.2.3.2 Markovian Stochastic State

Suppose now that the random process $\mathbf{V}_k$ is *white* noise, satisfying therefore the following relations

$$\mathrm{E}\left\{\mathbf{V}_k\right\} = 0 ,$$
$$\mathrm{Cov}\left\{\mathbf{V}_k, \mathbf{V}_j\right\} = \bar{\mathbf{Q}}_k\, \delta_{kj} .$$

Of course, this noise is correlated neither with the present system state nor with any past system state:

$$\mathrm{Cov}\left\{\mathbf{V}_k, \mathbf{X}_j\right\} = 0, \ \forall\, j \le k . \tag{4.13}$$

It is easy to see that in this case the conditional random variable $\mathbf{X}_{k+1} \mid \mathbf{X}_k$ does not depend on the past values $\mathbf{X}_j$, $\forall\, j < k$.

The state $\mathbf{X}_k$, at given *k*, is then a *Markovian* stochastic state, and the discrete stochastic process $\mathbf{X}_k$ defined by (4.12) is a *Markovian stochastic process*.

4.2.3.3 Solving the Linear Stochastic Difference Equation

The solution to (4.12) is thus as a Markovian stochastic process. Let us characterize this process.

The initial state $\mathbf{X}_0$ is itself random and is characterized by its mean $\boldsymbol{\mu}_{X,0}$ and its variance $\boldsymbol{\Sigma}_0$. The initial data of the problem are thus:

$$\mathrm{E}\left\{\mathbf{X}_0\right\} = \boldsymbol{\mu}_{X,0} \ \text{ and } \ \mathrm{Var}\left\{\mathbf{X}_0\right\} = \boldsymbol{\Sigma}_0 .$$

The process mean and its variance obey thus the following equations:

Mean value of the process:

$$\boldsymbol{\mu}_{X,\,k+1} = \mathrm{E}\left\{\mathbf{X}_{k+1}\right\} = \boldsymbol{\Phi}_k\,\boldsymbol{\mu}_{X,\,k}\,, \tag{4.14}$$

with given $\boldsymbol{\mu}_{X,0}$.

Consequence: the stochastic process $\mathbf{X}_k$ can always be assumed centered. Indeed,

– if $\boldsymbol{\mu}_{X,0} = 0$, we have from (4.14): $\boldsymbol{\mu}_{X,k} = 0,\ \forall k$;
– if $\boldsymbol{\mu}_{X,0} \neq 0$, it is possible to revert to the previous situation by a simple change of variable, by letting $\mathbf{Z}_k = \mathbf{X}_k - \boldsymbol{\mu}_{X,k}$, the process $\mathbf{Z}_k$ being then centered.

Autocovariance of the process: $\mathrm{Cov}\left\{\mathbf{X}_j, \mathbf{X}_k\right\} = \boldsymbol{\Sigma}_{j,k}$ **.**

With the assumption $\boldsymbol{\mu}_{X,k} = 0$, the autocovariance is expressed as

$$\boldsymbol{\Sigma}_{j,k} = \mathrm{E}\left\{\mathbf{X}_j\,\mathbf{X}_k^{\mathrm{T}}\right\}.$$

We calculate first $\boldsymbol{\Sigma}_{k+1,\,k}$:

$$\begin{aligned}\boldsymbol{\Sigma}_{k+1,\,k} = \mathrm{E}\left\{\mathbf{X}_{k+1}\,\mathbf{X}_k^{\mathrm{T}}\right\} &= \mathrm{E}\left\{(\boldsymbol{\Phi}_k\,\mathbf{X}_k + \mathbf{E}_k\mathbf{V}_k)\mathbf{X}_k^{\mathrm{T}}\right\}\\ &= \boldsymbol{\Phi}_k\,\mathrm{E}\left\{\mathbf{X}_k\,\mathbf{X}_k^{\mathrm{T}}\right\} + \mathbf{E}_k\,\mathrm{E}\left\{\mathbf{V}_k\,\mathbf{X}_k^{\mathrm{T}}\right\}.\end{aligned}$$

The second term of the right side vanishes according to (4.13), so that

$$\boldsymbol{\Sigma}_{k+1,\,k} = \boldsymbol{\Phi}_k\,\boldsymbol{\Sigma}_{k,\,k}\ \ . \tag{4.15}$$

Applying repetitively this relation, from step $k+1$ until step j ($j \geq k+1$), and using the fact that

$$\boldsymbol{\Phi}_{j-1}\boldsymbol{\Phi}_{j-2}\cdots\boldsymbol{\Phi}_k = \boldsymbol{\Phi}_{j,\,k}\,,$$

which is the basic property of the transition matrix (Sect. A.5), we get:

$$\boldsymbol{\Sigma}_{j,\,k} = \boldsymbol{\Phi}_{j,\,k}\,\boldsymbol{\Sigma}_{k,\,k},\ \text{ for } j \geq k\ \ . \tag{4.16}$$

This equality writes also

$$\boldsymbol{\Phi}_{j,\,k} = \boldsymbol{\Sigma}_{j,\,k}\,\boldsymbol{\Sigma}_{k,\,k}^{-1}\,. \tag{4.17}$$

It can be proved that this relation is characteristic of Markovian processes.

Variance of the process: $\mathrm{Var}\left\{\mathbf{X}_k\right\} = \mathbf{\Sigma}_k = \mathrm{E}\left\{\mathbf{X}_k \mathbf{X}_k^{\mathrm{T}}\right\}$.

Let us establish how this covariance matrix $\mathbf{\Sigma}_k$ of the stochastic state $\mathbf{X}_k$ is *propagated* from one step to the next, in absence of any measurement which might *improve* the knowledge of this state. According to (4.12), the following holds:

$$\begin{aligned}\mathbf{\Sigma}_{k+1} &= \mathrm{E}\left\{\mathbf{X}_{k+1}\mathbf{X}_{k+1}^{\mathrm{T}}\right\} = \mathrm{E}\left\{(\mathbf{\Phi}_k \mathbf{X}_k + \mathbf{E}_k\mathbf{V}_k)(\mathbf{\Phi}_k \mathbf{X}_k + \mathbf{E}_k\mathbf{V}_k)^{\mathrm{T}}\right\} \\ &= \mathrm{E}\left\{\mathbf{\Phi}_k \mathbf{X}_k\mathbf{X}_k^{\mathrm{T}} \mathbf{\Phi}_k^{\mathrm{T}} + \mathbf{E}_k\mathbf{V}_k\mathbf{X}_k^{\mathrm{T}} \mathbf{\Phi}_k^{\mathrm{T}} + \mathbf{\Phi}_k \mathbf{X}_k\mathbf{V}_k^{\mathrm{T}}\mathbf{E}_k^{\mathrm{T}} + \mathbf{E}_k\mathbf{V}_k\mathbf{V}_k^{\mathrm{T}}\mathbf{E}_k^{\mathrm{T}}\right\}.\end{aligned}$$

Due to (4.13), this expression simplifies to

$$\mathbf{\Sigma}_{k+1} = \mathbf{\Phi}_k \mathbf{\Sigma}_k \mathbf{\Phi}_k^{\mathrm{T}} + \mathbf{E}_k \bar{\mathbf{Q}}_k \mathbf{E}_k^{\mathrm{T}} \ . \tag{4.18}$$

This equation is called the *propagation equation of the variance* or of the *covariance matrix*. It has the initial condition: $\mathbf{\Sigma}_0 = \mathrm{Var}\left\{\mathbf{X}_0\right\}$.

4.2.3.4 State Model of a Discrete Markovian Process

We can now inverse the problem, and consider that a given Markovian process is built by feeding white noise at the input of a linear system, to be determined.

From (4.15) and (4.18) the following theorem is deduced:

Theorem 4.2. *Every discrete Markovian stochastic process* $\mathbf{X}_k$*, of mean* $\boldsymbol{\mu}_{X,k}$ *and of covariance* $\mathbf{\Sigma}_{j,k}$*, can be represented by the state of a discrete linear system fed by a discrete white noise.*
This model is described by the linear stochastic equation

$$\mathbf{X}_{k+1} = \mathbf{\Phi}_k \mathbf{X}_k + \mathbf{E}_k \mathbf{V}_k \, ,$$

with:

$$\mathrm{E}\left\{\mathbf{X}_0\right\} = \boldsymbol{\mu}_{X,0} \, , \quad \mathrm{Var}\left\{\mathbf{X}_0\right\} = \mathbf{\Sigma}_0 \, ,$$

$$\mathrm{E}\left\{\mathbf{V}_k\right\} = 0 \, ,$$

$$\mathrm{Cov}\left\{\mathbf{V}_k, \mathbf{V}_j\right\} = \bar{\mathbf{Q}}_k \, \delta_{kj} \, ,$$

$$\mathrm{Cov}\left\{\mathbf{V}_k, \mathbf{X}_j\right\} = 0 \quad \forall \, j \le k \, ,$$

and where

$$\mathbf{\Phi}_k = \mathbf{\Sigma}_{k+1,k}\, \mathbf{\Sigma}_{k,k}^{-1} ,$$

$$\begin{aligned}\mathbf{E}_k \bar{\mathbf{Q}}_k \mathbf{E}_k^{\mathrm{T}} &= \mathbf{\Sigma}_{k+1} - \mathbf{\Phi}_k\, \mathbf{\Sigma}_k\, \mathbf{\Phi}_k^{\mathrm{T}} \\ &= \mathbf{\Sigma}_{k+1} - \mathbf{\Sigma}_{k+1,\,k}\, \mathbf{\Sigma}_{k,\,k}^{-1}\, \mathbf{\Sigma}_{k,\,k+1} .\end{aligned}$$

4.2.3.5 Generalized Discrete Model

The previously considered system was a linear system fed by white noise. This noise $\mathbf{V}_k$ affected the state $\mathbf{X}_k$ of the system, which from deterministic had become random. For this reason, $\mathbf{V}_k$ is called *state noise* or *process noise*.

Let us complete this model by taking now into account the output $\mathbf{y}_k$ of such a system, itself corrupted by a *measurement noise* $\mathbf{W}_k$.

Consider the linear system, governed by the following state equations:

$$\begin{cases}\mathbf{X}_{k+1} = \mathbf{\Phi}_k\, \mathbf{X}_k + \mathbf{E}_k \mathbf{V}_k \\ \quad \mathbf{Y}_k = \mathbf{C}_k\, \mathbf{X}_k + \mathbf{W}_k\end{cases} \tag{4.19}$$

where $\mathbf{\Phi}_k$, $\mathbf{E}_k$ and $\mathbf{V}_k$ are defined as previously.

The measurement $\mathbf{Y}_k$ consists of a known linear function of the state, $\mathbf{C}_k \mathbf{X}_k$, perturbed by the discrete noise $\mathbf{W}_k$, which we assume again to be white noise. It has thus become also a random vector.

The state $\mathbf{X}_k$ and the measurement noise $\mathbf{W}_k$ are uncorrelated stochastic processes, with the following means and covariances:

$$\mathrm{E}\{\mathbf{X}_k\} = \boldsymbol{\mu}_{X,k} , \quad \mathrm{Cov}\{\mathbf{X}_k, \mathbf{X}_j\} = \mathbf{\Sigma}_{k,j} , \tag{4.20}$$

$$\mathrm{E}\{\mathbf{W}_k\} = 0 , \quad \mathrm{Cov}\{\mathbf{W}_k, \mathbf{W}_j\} = \bar{\mathbf{R}}_k\, \delta_{kj} , \tag{4.21}$$

$$\mathrm{Cov}\{\mathbf{W}_k, \mathbf{X}_j\} = 0, \quad \forall\, j,k . \tag{4.22}$$

The system described by (4.19) is represented in Fig. 4.2.

$\mathbf{Y}_k$ defined this way represents still a discrete stochastic, but *no longer Markovian*, process with the two following moments:

$$\boldsymbol{\mu}_{Y,k} = \mathbf{C}_k\, \boldsymbol{\mu}_{X,k} , \tag{4.23}$$

$$\mathrm{Cov}\{\mathbf{Y}_k, \mathbf{Y}_j\} = \mathbf{C}_k\, \mathbf{\Sigma}_{k,j}\, \mathbf{C}_j^{\mathrm{T}} + \bar{\mathbf{R}}_k\, \delta_{kj} . \tag{4.24}$$

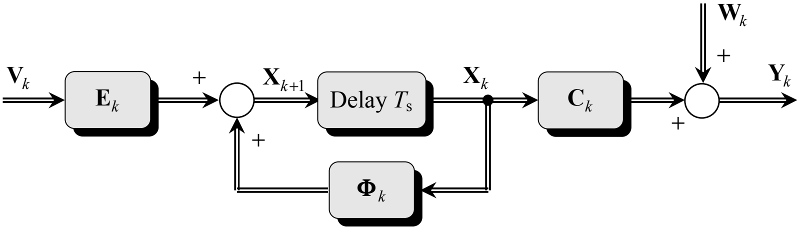


Fig. 4.2 Discrete stochastic system with process and measurement noise.

The proof of (4.23) is trivial, the noise $\mathbf{W}_k$ being centered. To prove (4.24), we write:

$$\begin{aligned}\operatorname{Cov}\left\{\mathbf{Y}_k,\mathbf{Y}_j\right\} &= \mathrm{E}\left\{(\mathbf{Y}_k-\boldsymbol{\mu}_{Y,k})(\mathbf{Y}_j-\boldsymbol{\mu}_{Y,j})^{\mathrm{T}}\right\}\\ &= \mathrm{E}\left\{\left[\mathbf{C}_k(\mathbf{X}_k-\boldsymbol{\mu}_{X,k})+\mathbf{W}_k\right]\left[\mathbf{C}_j(\mathbf{X}_j-\boldsymbol{\mu}_{X,j})+\mathbf{W}_j\right]^{\mathrm{T}}\right\}\\ &= \mathbf{C}_k\,\mathrm{E}\left\{(\mathbf{X}_k-\boldsymbol{\mu}_{X,k})(\mathbf{X}_j-\boldsymbol{\mu}_{X,j})^{\mathrm{T}}\right\}\mathbf{C}_j^{\mathrm{T}}+\mathbf{C}_k\,\mathrm{E}\left\{(\mathbf{X}_k-\boldsymbol{\mu}_{X,k})\mathbf{W}_j^{\mathrm{T}}\right\}\\ &\qquad +\mathrm{E}\left\{\mathbf{W}_k(\mathbf{X}_j-\boldsymbol{\mu}_{X,j})^{\mathrm{T}}\right\}\mathbf{C}_j^{\mathrm{T}}+\mathrm{E}\left\{\mathbf{W}_k\mathbf{W}_j^{\mathrm{T}}\right\}.\end{aligned}$$

Since $\mathbf{W}_k$ is centered, the second and third term on the right side of this equation amount respectively to $\mathbf{C}_k\operatorname{Cov}\left\{\mathbf{X}_k,\mathbf{W}_j\right\}$ and $\operatorname{Cov}\left\{\mathbf{W}_k,\mathbf{X}_j\right\}\mathbf{C}_j^{\mathrm{T}}$, and vanish thus because of (4.22). The proof is then completed simply by the use of (4.20) and of (4.21) for the two remaining terms.

4.2.3.6 Case of Stationary Processes

If we start from a stable discrete linear and *time-invariant* system, thus described by constant matrices $\mathbf{\Phi}$, $\mathbf{E}$ and $\mathbf{C}$:

$$\begin{cases}\mathbf{X}_{k+1}=\mathbf{\Phi}\mathbf{X}_k+\mathbf{E}\,\mathbf{V}_k\\ \mathbf{Y}_k=\mathbf{C}\,\mathbf{X}_k+\mathbf{W}_k\end{cases}$$

and if we suppose the white noises $\mathbf{V}_k$ and $\mathbf{W}_k$ *stationary,* i.e.:

$$\bar{\mathbf{Q}}_k=\bar{\mathbf{Q}},\quad \bar{\mathbf{R}}_k=\bar{\mathbf{R}},$$

where $\bar{\mathbf{Q}}$ and $\bar{\mathbf{R}}$ are constant matrices, the equations of the previous sections are modified as follows:

Transition matrix:

$$\mathbf{\Phi}_{j,k} = \mathbf{\Phi}(j-k) = \mathbf{\Phi}(m),$$

where $m = j - k$. The transition matrix depends thus only on the time interval, expressed in discrete time steps. Recall that, for $m = 1$, $\mathbf{\Phi}_k = \mathbf{\Phi}$, and that, in the case of a discrete-time linear system resulting from the sampling at a constant period T_s of a continuous-time linear system with system matrix **A**, this transition matrix amounts to $\mathbf{\Phi} = \exp(\mathbf{A}T_s)$, as in (A.23) in Appendix A.

Consequences:

1. The stochastic process $\mathbf{X}_k$ is *stationary.*

   Indeed, according to (4.16), $\mathbf{\Sigma}_{j,k}$ is then, itself also, only function of $j-k$:

   $$\mathbf{\Sigma}_{j,k} = \mathbf{\Sigma}(j-k),$$

   which implies that

   $$\mathbf{\Sigma}(k,k) = \mathbf{\Sigma}_{k,k} = \mathbf{\Sigma}(0) = \mathbf{\Sigma},$$

   which is a constant matrix.

2. Covariance matrix:

   (4.15) is therefore replaced by

   $$\mathbf{\Sigma}_{k+1,k} = \mathbf{\Phi}\,\mathbf{\Sigma}, \tag{4.25}$$

   whereas (4.16) becomes:

   $$\mathbf{\Sigma}_{k+m,k} = \mathrm{E}\left\{\mathbf{X}_{k+m}\,\mathbf{X}_k^{\mathrm{T}}\right\} = \begin{cases} \mathbf{\Phi}^m\mathbf{\Sigma}, & \text{if } m \geq 0 \\ \mathbf{\Sigma}(\mathbf{\Phi}^{\mathrm{T}})^{-m} & \text{if } m \leq 0 \end{cases} \tag{4.26}$$

   Proof:
   if $m \geq 0$, by repetitive application of (4.25);
   if $m \leq 0$: by transposition of the equation corresponding to $m \geq 0$, while letting $p = -m \geq 0$, we have

   $$\mathbf{\Sigma}_{k-p,k} = \mathbf{\Sigma}_{k,k-p}^{\mathrm{T}} = (\mathbf{\Phi}^p\,\mathbf{\Sigma})^{\mathrm{T}} = \mathbf{\Sigma}(\mathbf{\Phi}^{\mathrm{T}})^p.$$

3. Variance:

As to (4.18), it is replaced by

$$\mathbf{\Sigma} = \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^{\mathrm{T}} + \mathbf{E}\bar{\mathbf{Q}}\mathbf{E}^{\mathrm{T}} . \tag{4.27}$$

The rewriting of Theorem 4.2 for the present case is derived easily from these relations, and the same holds for the equations replacing (4.23) and (4.24).

4.2.4 Continuous Markovian Processes. Noisy Continuous Systems

4.2.4.1 Stochastic State Model. Stochastic Differential Equation

Let us consider a continuous-time linear system, not necessarily time-invariant in the most general case, and fed by a continuous-time white noise. Its state equation becomes then a *stochastic linear differential equation*:

$$\dot{\mathbf{X}}(t) = \mathbf{A}(t)\mathbf{X}(t) + \mathbf{E}(t)\mathbf{V}(t) \ , \tag{4.28}$$

where $\mathbf{V}(t)$ denotes a continuous-time white noise, uncorrelated with the system past state $\mathbf{X}(t)$, thus satisfying:

$$\mathrm{E}\{\mathbf{V}(t)\} = 0 ,$$

$$\mathrm{Cov}\{\mathbf{V}(t), \mathbf{V}(t')\} = \bar{\mathbf{Q}}(t)\,\delta(t - t') ,$$

$$\mathrm{Cov}\{\mathbf{V}(t), \mathbf{X}(t')\} = 0 \quad \forall\, t' < t .$$

Such a system is represented in Fig. 4.3.

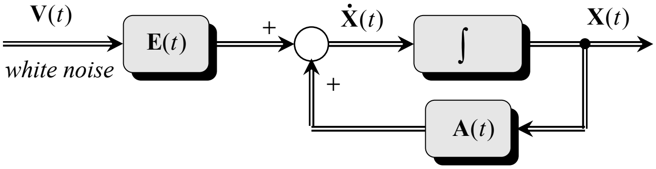


Fig. 4.3 Time-varying, linear, continuous-time system with random input and random state.

4.2.4.2 Passage from Discrete to Continuous Case

Rather than reestablishing the formulae corresponding to the case of a continuous system, we will derive them from the previous study by considering the continuous system as the limit of a discrete system whose sampling time T_s would tend towards an infinitesimally small value:

$$(k+1)T_s - kT_s = T_s = \Delta t \to dt\ .$$

Relations governing the passage to the limit.

- State equations:

$$\mathbf{X}_{k+1} \longrightarrow \mathbf{X}(t) + d\mathbf{X} \tag{4.29}$$

$$\mathbf{\Phi}_k \longrightarrow \mathbf{I} + \mathbf{A}(t)\,dt \tag{4.30}$$

$$\mathbf{E}_k \longrightarrow \mathbf{E}(t)\,dt \tag{4.31}$$

$$\mathbf{V}_k \longrightarrow \mathbf{V}(t) \tag{4.32}$$

- Noise covariance matrix:

$$\bar{\mathbf{Q}}_k \longrightarrow \frac{\bar{\mathbf{Q}}(t)}{dt}\ . \tag{4.33}$$

Proof. Let us write side by side the state equations for the two system classes:

$$\begin{array}{r|l} & \dot{\mathbf{X}}(t) = \dfrac{d\mathbf{X}(t)}{dt} = \mathbf{A}(t)\mathbf{X}(t) + \mathbf{E}(t)\mathbf{V}(t) \\ & d\mathbf{X}(t) = \mathbf{A}(t)\mathbf{X}(t)\,dt + \mathbf{E}(t)\mathbf{V}(t)\,dt \\ \mathbf{X}_{k+1} = \mathbf{\Phi}_k\mathbf{X}_k + \mathbf{E}_k\mathbf{V}_k & \mathbf{X}(t) + d\mathbf{X}(t) = \left[\mathbf{I} + \mathbf{A}(t)dt\right]\mathbf{X}(t) + \mathbf{E}(t)dt\,\mathbf{V}(t) \end{array}$$

Relations (4.29) to (4.32) result readily from the simple comparison of the two cases, $\mathbf{X}_k$ corresponding to $\mathbf{X}(t)$ and $\mathbf{X}_{k+1}$ to $\mathbf{X}(t) + d\mathbf{X}(t)$.

The case of the passage relation (4.33) is more delicate to handle. What we are looking for is the equivalence between the discrete sequence of white noise $\mathbf{V}_k$ and the continuous process of white noise $\mathbf{V}(t)$, which is not physically realizable.

Remark first a difference in dimensions between the two $\bar{\mathbf{Q}}$ matrices. Indeed, the covariance matrices of the noise are, respectively:

- in the discrete case: $\bar{\mathbf{Q}}_k\,\delta_{kj}$;
- in the continuous case: $\bar{\mathbf{Q}}(t)\,\delta(t-t')$.

$\bar{\mathbf{Q}}_k$ is therefore a covariance matrix. But, since the Dirac distribution $\delta(t-t')$ has the dimension of the inverse of time, as shows its basic property $\int_{-\infty}^{\infty} \delta(t)\,dt = 1$, the matrix $\bar{\mathbf{Q}}(t)$ is, on the contrary, a power spectral density matrix. Recall indeed that, in the case of a white noise, a power spectral density matrix can be converted to a covariance matrix by multiplying it with a Dirac distribution, as shown in the stationary case by (C.19) of Appendix C, by letting there $\mathbf{\Lambda}_{XX}(\tau) = \mathbf{\Sigma}_{XX}(\tau)$ (centered noise) and $\mathbf{\Phi}_{XX}(\omega) = \mathbf{\Phi}_{XX}$ (constant matrix, the noise being white):

$$\mathbf{\Sigma}_{XX}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathbf{\Phi}_{XX}(\omega)\, e^{j\omega\tau} d\omega = \mathbf{\Phi}_{XX}\,\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{j\omega\tau} d\omega = \mathbf{\Phi}_{XX}\,\delta(\tau)\,,$$

where use has been made of the inverse Fourier transform of 1.

The covariance matrix $\bar{\mathbf{Q}}(t)\delta(t-t')$ has thus infinite elements. To obtain the continuous white noise process as a limit, for T_s becoming infinitesimally small (*dt*), of a sequence of discrete white noise, it is necessary to increase the amplitude of the discrete impulses at the same time that their duration T_s is decreased, while maintaining constant the power spectral density. Let us clarify this point.

Since white noise is a centered process, its autocorrelation matrix is equal to its covariance matrix:

- discrete case: $\bar{\mathbf{Q}}_k\ \delta_{kj} = \mathrm{E}\left\{\mathbf{V}_k \mathbf{V}_j^{\mathrm{T}}\right\}$;
- continuous case: $\bar{\mathbf{Q}}(t)\ \delta(t-t') = \mathrm{E}\left\{\mathbf{V}(t)\mathbf{V}^{\mathrm{T}}(t')\right\} = \mathbf{\Sigma}_{VV}(t,t')$.

To realize the equality of the power spectral densities in these two cases means, since in both cases we deal with white noises, thus having constant spectral densities, to realize their equality for $\omega = 0$. According to (C.18), Appendix C, this amounts to equalizing the following integrals of their autocorrelation matrices:

- discrete case: $\int_{-\infty}^{\infty} \bar{\mathbf{Q}}_k\ \delta_{kj} dt = \int_{kT_s}^{(k+1)T_s} \bar{\mathbf{Q}}_k\ dt = \bar{\mathbf{Q}}_k\ T_s$;
- continuous case: $\int_{-\infty}^{\infty} \bar{\mathbf{Q}}(t)\ \delta(t-t')dt = \bar{\mathbf{Q}}(t')$.

The equality of these two expressions at the limit $T_s \rightarrow dt$ imposes the correspondence

$$\bar{\mathbf{Q}}_k dt \longrightarrow \bar{\mathbf{Q}}(t)\,.$$

which is indeed (4.33).

4.2.4.3 Solving the Stochastic Linear Differential Equation

Let us now solve (4.28). Since the continuous white noise $\mathbf{V}(t)$ is not correlated with the past state $\mathbf{X}(t)$, it is easy to verify that here again the conditional random variable $\mathbf{X}(u)\,|\,\mathbf{X}(t)$, where $t<u$, does not depend on the past values $\mathbf{X}(t_1)$, $\forall\, t_1 < t$.

The solution of (4.28) is therefore a *Markovian* continuous-time stochastic process $\mathbf{X}(t)$. The initial state $\mathbf{X}_0$ is again assumed to have the mean $\boldsymbol{\mu}_{X,0}$ and the variance $\boldsymbol{\Sigma}_0$.

Let us derive now from the discrete case, by means of the passage to the limit formulae (4.29) to (4.33), the two first moments of the process $\mathbf{X}(t)$.

Mean value of the process $\mathbf{X}(t)$:

From (4.14) we deduce by applying the first passage formula that

$$\boldsymbol{\mu}_X(t) + d\,\boldsymbol{\mu}_X(t) = \left[\mathbf{I} + \mathbf{A}(t)dt\right]\boldsymbol{\mu}_X(t),$$

i.e.

$$\dot{\boldsymbol{\mu}}_X(t) = \mathbf{A}(t)\,\boldsymbol{\mu}_X(t).$$

This equation has the same form as the one governing the free response of a linear system. Its solution is given by (Sect. A.5, Appendix A):

$$\boldsymbol{\mu}_X(t) = \boldsymbol{\Phi}(t,0)\,\boldsymbol{\mu}_{X,0}. \qquad (4.34)$$

Covariance of the process $\mathbf{X}(t)$:

We will calculate it from (4.16), where the discrete times t_j, t_k are simply replaced by the times t_1, t_2 belonging to a continuous time scale. Hence

$$\boldsymbol{\Sigma}(t_2,t_1) = \boldsymbol{\Phi}(t_2,t_1)\,\boldsymbol{\Sigma}(t_1,t_1), \ \text{ for } t_2 \ge t_1, \qquad (4.35)$$

which can also be written

$$\boldsymbol{\Phi}(t_2,t_1) = \boldsymbol{\Sigma}(t_2,t_1)\,\boldsymbol{\Sigma}^{-1}(t_1,t_1), \ \text{ for } t_2 \ge t_1, \qquad (4.36)$$

if $\boldsymbol{\Sigma}(t,t)$ is invertible.

Variance of the process $\mathbf{X}(t)$:

The variance propagation equation of the discrete case, (4.18), is transformed by applying the relations of passage to the limit of Sect. 4.2.4.2 and by omitting momentarily the time dependence, in order to simplify the writing:

$$\begin{aligned}\mathbf{\Sigma} + d\mathbf{\Sigma} &= (\mathbf{I} + \mathbf{A}\,dt)\,\mathbf{\Sigma}\,(\mathbf{I} + \mathbf{A}\,dt)^{\mathrm{T}} + \mathbf{E}\,dt\frac{\bar{\mathbf{Q}}}{dt}\mathbf{E}^{\mathrm{T}}dt \\ &= \mathbf{\Sigma} + \mathbf{A}\mathbf{\Sigma}\,dt + \mathbf{\Sigma}\mathbf{A}^{\mathrm{T}}dt + \mathbf{A}\mathbf{\Sigma}\mathbf{A}^{\mathrm{T}}(dt)^2 + \mathbf{E}\bar{\mathbf{Q}}\mathbf{E}^{\mathrm{T}}dt\,.\end{aligned}$$

Neglecting the infinitesimally small term of second order, we obtain:

$$d\mathbf{\Sigma} = \mathbf{A}\mathbf{\Sigma}\,dt + \mathbf{\Sigma}\mathbf{A}^{\mathrm{T}}dt + \mathbf{E}\bar{\mathbf{Q}}\mathbf{E}^{\mathrm{T}}dt\,,$$

i.e. also:

$$\dot{\mathbf{\Sigma}}(t) = \mathbf{A}(t)\,\mathbf{\Sigma}(t) + \mathbf{\Sigma}(t)\,\mathbf{A}^{\mathrm{T}}(t) + \mathbf{E}(t)\,\bar{\mathbf{Q}}(t)\,\mathbf{E}^{\mathrm{T}}(t)\ . \tag{4.37}$$

This equation is called *linear evolution equation of the variance*, or of the *covariance matrix*. It is solved from the initial condition $\mathbf{\Sigma}(0) = \mathbf{\Sigma}_0$.

Particular case. If $\mathbf{V}(t)$ and $\mathbf{X}_0$ are Gaussian, the solution $\mathbf{X}(t)$ of (4.28) represents a Gaussian-Markovian process.

4.2.4.4 State Model of a Continuous Markovian Stochastic Process

An analog approach to that of the discrete case leads to the following statement:

Theorem 4.3. *Every continuous-time Markovian stochastic process,* $\mathbf{X}(t)$*, with mean* $\boldsymbol{\mu}_X(t)$ *and with covariance* $\mathbf{\Sigma}(t_2, t_1)$*, can be represented as the state of a continuous linear system fed by a continuous white noise.*
The linear system is described by the equation

$$\dot{\mathbf{X}}(t) = \mathbf{A}(t)\mathbf{X}(t) + \mathbf{E}(t)\mathbf{V}(t)\,,$$

where $\mathbf{A}(t)$ *is defined by*

$$\dot{\mathbf{\Phi}}(t_2,t_1) = \mathbf{A}(t_2)\,\mathbf{\Phi}(t_2,t_1)\,,$$

with $$\mathbf{\Phi}(t_2,t_1) = \mathbf{\Sigma}(t_2,t_1)\,\mathbf{\Sigma}^{-1}(t_1,t_1), \quad t_1 \le t_2\,,$$

and where $\mathbf{V}(t)$ *is defined by*

$$\begin{aligned}&\mathrm{E}\{\mathbf{V}(t)\} = 0\,,\\ &\mathrm{Cov}\{\mathbf{V}(t),\mathbf{V}(t')\} = \bar{\mathbf{Q}}(t)\,\delta(t-t')\,,\\ &\mathbf{E}(t)\,\bar{\mathbf{Q}}(t)\,\mathbf{E}^{\mathrm{T}}(t) = \dot{\mathbf{\Sigma}}(t) - \mathbf{A}(t)\mathbf{\Sigma}(t) - \mathbf{\Sigma}(t)\mathbf{A}^{\mathrm{T}}(t)\,.\end{aligned}$$

4.2.4.5 Generalized Continuous Model

The previous model is completed by the addition of a measurement equation:

$$\begin{cases} \dot{\mathbf{X}}(t) = \mathbf{A}(t)\mathbf{X}(t) + \mathbf{E}(t)\mathbf{V}(t) \\ \mathbf{Y}(t) = \mathbf{C}(t)\mathbf{X}(t) + \mathbf{W}(t) \end{cases} \tag{4.38}$$

with the following supplementary assumptions:

$$\mathrm{E}\{\mathbf{X}(t)\} = \boldsymbol{\mu}_X(t)\,; \quad \mathrm{Cov}\{\mathbf{X}(t), \mathbf{X}(t')\} = \boldsymbol{\Sigma}(t,t')\,,$$

$$\mathrm{E}\{\mathbf{W}(t)\} = 0\,; \quad \mathrm{Cov}\{\mathbf{W}(t), \mathbf{W}(t)'\} = \overline{\mathbf{R}}(t)\,\delta(t-t')\,,$$

$$\mathrm{Cov}\{\mathbf{X}(t), \mathbf{W}(t')\} = 0, \quad \forall\, t,t'\,.$$

The system described by (4.38) is represented in Fig. 4.4.

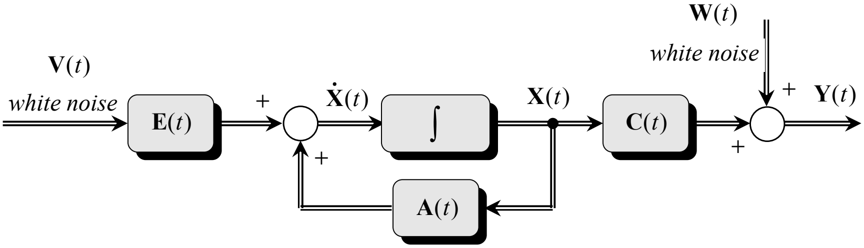


Fig. 4.4 Time-varying continuous-time stochastic system with process and measurement noise.

The stochastic process $\mathbf{Y}(t)$ produced this way *is no longer Markovian*. It has the two following moments:

$$\boldsymbol{\mu}_Y(t) = \mathbf{C}(t)\,\boldsymbol{\mu}_X(t)\,, \tag{4.39}$$

$$\mathrm{Cov}\{\mathbf{Y}(t), \mathbf{Y}(t')\} = \mathbf{C}(t)\boldsymbol{\Sigma}(t,t')\mathbf{C}^{\mathrm{T}}(t') + \overline{\mathbf{R}}(t)\delta(t-t')\,. \tag{4.40}$$

The proof of the second of these formulae is very similar to the one which has been given in the discrete case, equation (4.24).

4.2.4.6 Case of Stationary Processes

Again, the solution in case of a time-invariant system will be a stationary process:

$$\begin{cases}\dot{\mathbf{X}} = \mathbf{AX} + \mathbf{EV} \\ \mathbf{Y} = \mathbf{CX} + \mathbf{W}\end{cases}$$

where the white noises $\mathbf{V}(t)$ and $\mathbf{W}(t)$ are stationary, i.e. $\overline{\mathbf{Q}}(t) = \overline{\mathbf{Q}}$ constant and $\overline{\mathbf{R}}(t) = \overline{\mathbf{R}}$ constant. The previous theory is modified as follows:

Transition matrix:

$$\mathbf{\Phi}(t_2, t_1) = \mathbf{\Phi}(t_2 - t_1) = \mathbf{\Phi}(\tau) = e^{\mathbf{A}\tau} .$$

Variance:

$$\mathbf{\Sigma}(t,t) = \mathbf{\Sigma} ,$$

where $\mathbf{\Sigma}$ is a constant matrix. Equation (4.37) is replaced by

$$\mathbf{A\Sigma} + \mathbf{\Sigma A}^{\mathrm{T}} + \mathbf{E}\overline{\mathbf{Q}}\mathbf{E}^{\mathrm{T}} = 0 . \tag{4.41}$$

Covariance: The equation (4.35) becomes now

$$\begin{aligned} &\mathbf{\Sigma}(\tau) = e^{\mathbf{A}\tau}\mathbf{\Sigma} && \text{if } \tau > 0 \\ &\mathbf{\Sigma}(\tau) = \mathbf{\Sigma} e^{-\mathbf{A}^{\mathrm{T}}\tau} && \text{if } \tau < 0 . \end{aligned} \tag{4.42}$$

Proof of the second formula. If $\tau < 0$, according to the first of these formulae, $\mathbf{\Sigma}(-\tau) = e^{-\mathbf{A}\tau}\mathbf{\Sigma}$, then, according to (C.17), Appendix C:

$$\mathbf{\Sigma}(\tau) = (e^{-\mathbf{A}\tau}\mathbf{\Sigma})^{\mathrm{T}} = \mathbf{\Sigma}^{\mathrm{T}}(e^{-\mathbf{A}\tau})^{\mathrm{T}} = \mathbf{\Sigma}^{\mathrm{T}} e^{-\mathbf{A}^{\mathrm{T}}\tau} = \mathbf{\Sigma}\, e^{-\mathbf{A}^{\mathrm{T}}\tau} ,$$

since a covariance matrix is always symmetric ($\mathbf{\Sigma}^{\mathrm{T}} = \mathbf{\Sigma}$).

4.3 Optimal Linear Filtering

4.3.1 Discrete Kalman Filter

As mentioned in the introduction to the present chapter, we will establish first the equations of the discrete Kalman filter, which is easier to present mainly for notational reasons.

4.3.1.1 Presentation of the problem

We want to estimate the stochastic state vector $\mathbf{X}$, of dimension n, of a discrete linear system submitted to disturbances (process noise $\mathbf{V}$), by means of the measurement vector $\mathbf{Y}$, of dimension q, in the presence of a measurement noise $\mathbf{W}$ independent from the state $\mathbf{X}$.

The state space representation of a linear system affected by such noise sources is thus the following:

$$\begin{cases} \mathbf{X}_{k+1} = \mathbf{\Phi}_k \, \mathbf{X}_k + \mathbf{E}_k \mathbf{V}_k \\ \mathbf{Y}_k = \mathbf{C}_k \, \mathbf{X}_k + \mathbf{W}_k \end{cases} \tag{4.43}$$

where it is assumed that the process noise $\mathbf{V}_k$ and the measurement noise $\mathbf{W}_k$ are distinct white noise sources, thus not correlated with each other. The assumptions concerning these equations can thus be summarized as follows:

$$\mathrm{E}\left\{\mathbf{V}_k\right\} = 0 \,, \; \mathrm{Cov}\left\{\mathbf{V}_k, \mathbf{V}_j\right\} = \bar{\mathbf{Q}}_k \, \delta_{kj} \,,$$

$$\mathrm{Cov}\left\{\mathbf{V}_k, \mathbf{X}_j\right\} = 0, \quad \forall \, j \le k \,,$$

$$\mathrm{E}\left\{\mathbf{X}_0\right\} = \boldsymbol{\mu}_{X,0} \,, \; \mathrm{Var}\left\{\mathbf{X}_0\right\} = \mathbf{\Sigma}_0 \,,$$

$$\mathrm{Var}\left\{\mathbf{X}_k\right\} = \mathbf{\Sigma}_k \,,$$

$$\mathrm{E}\left\{\mathbf{W}_k\right\} = 0 \,, \; \mathrm{Cov}\left\{\mathbf{W}_k, \mathbf{W}_j\right\} = \bar{\mathbf{R}}_k \, \delta_{kj} \,,$$

$$\mathrm{Cov}\left\{\mathbf{W}_k, \mathbf{X}_j\right\} = 0, \quad \forall \, j,k \,. \tag{4.44}$$

By definition, $\bar{\mathbf{Q}}_k$ is a symmetric, positive semidefinite matrix (Appendix C, Sect. C.2.6). We assume in addition that $\bar{\mathbf{R}}_k$, also symmetric, is positive definite, i.e. that the various measurement noises are statistically independent, which is a sound hypothesis if the measurement sensors are distinct. We assume further that the pair $(\mathbf{\Phi}_k, \mathbf{C}_k)$ is detectable $\forall k$, i.e. that the model contains no unstable *and* unobservable eigenvalue.

This system will thus be represented by the generalized discrete model, illustrated in Fig. 4.2.

The present objective is to determine the best estimate of the state $\mathbf{X}_k$, in the sense of the minimal error variance, by using only the measurement sequence $\mathbf{y}_0, \mathbf{y}_1, \cdots, \mathbf{y}_{k-1}, \mathbf{y}_k$.

Define:

- $\hat{\mathbf{x}}_{k|k}$ as the optimal estimate of $\mathbf{X}_k$ elaborated *after* the measurement $\mathbf{y}_k$, thus by taking into account all the measurements of the above sequence; we will call it in the sequel *estimation* of the state $\mathbf{X}_k$;
- $\hat{\mathbf{x}}_{k|k-1}$ as the optimal estimate of the same state $\mathbf{X}_k$ elaborated *before* the measurement $\mathbf{y}_k$, i.e. from the measurement sequence $\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_{k-1}$; we will call this estimate *prediction* of the state $\mathbf{X}_k$.

4.3.1.2 Estimation and Prediction Equations

In order to establish the discrete Kalman filter equations, we will proceed initially in three phases, which will merge afterwards into only two phases.

1st phase. Consider the system at step $k-1$.
$\mathbf{X}_{k-1}$ and $\mathbf{V}_{k-1}$ being random vectors, the state vector $\mathbf{X}_k$ of the next step will be random also. The best estimation of it that it is possible to make at this stage of knowledge, $\hat{\mathbf{x}}_{k|k-1}$, is its mean value or expectation

$$\hat{\mathbf{x}}_{k|k-1} = \mathrm{E}\left\{\mathbf{X}_k\right\} = \boldsymbol{\mu}_{X,k} .$$

The estimation error being given by $\tilde{\mathbf{X}}_{k|k-1} = \mathbf{X}_k - \hat{\mathbf{x}}_{k|k-1} = \mathbf{X}_k - \boldsymbol{\mu}_{X,k}$, it is evident that this estimation is unbiased $\mathrm{E}\left\{\tilde{\mathbf{X}}_{k|k-1}\right\} = \boldsymbol{\mu}_{X,k} - \boldsymbol{\mu}_{X,k} = 0$. Its variance amounts thus to

$$\mathrm{Var}\left\{\tilde{\mathbf{X}}_{k|k-1}\right\} = \mathrm{E}\left\{\tilde{\mathbf{X}}_{k|k-1}\,\tilde{\mathbf{X}}_{k|k-1}^{\mathrm{T}}\right\} = \mathrm{E}\left\{\left[\mathbf{X}_k - \boldsymbol{\mu}_{X,k}\right]\left[\mathbf{X}_k - \boldsymbol{\mu}_{X,k}\right]^{\mathrm{T}}\right\} = \mathrm{Var}\left\{\mathbf{X}_k\right\} .$$

$\tilde{\mathbf{X}}_{k|k-1}$ being the estimation error of $\mathbf{X}_k$ *before* the measurement $\mathbf{y}_k$, i.e. from the measurement sequence $\left\{\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_{k-1}\right\}$, its variance $\mathrm{Var}\left\{\tilde{\mathbf{X}}_{k|k-1}\right\}$ will be denoted by $\boldsymbol{\Sigma}_{k|k-1}$, using the same notation as the one adopted for the estimates.

According to (4.14) and (4.18), the previous values are given respectively by

$$\boldsymbol{\mu}_{X,k} = \boldsymbol{\Phi}_{k-1}\,\boldsymbol{\mu}_{X,k-1} ,$$

which can also be written

$$\hat{\mathbf{x}}_{k|k-1} = \boldsymbol{\Phi}_{k-1}\,\hat{\mathbf{x}}_{k-1|k-2} , \tag{4.45}$$

and by

$$\boldsymbol{\Sigma}_{k|k-1} = \boldsymbol{\Phi}_{k-1}\,\boldsymbol{\Sigma}_{k-1|k-2}\,\boldsymbol{\Phi}_{k-1}^{\mathrm{T}} + \mathbf{E}_{k-1}\,\overline{\mathbf{Q}}_{k-1}\,\mathbf{E}_{k-1}^{\mathrm{T}}\,. \tag{4.46}$$

2nd phase. Suppose now that the transition at step k takes place, and that a measurement occurs *after* the transition. This measurement yields a realization $\mathbf{y}_k$ of $\mathbf{Y}_k = \mathbf{C}_k\,\mathbf{X}_k + \mathbf{W}_k$, with the assumptions given at the beginning of this section. According to (4.4), the best estimate of $\mathbf{X}_k$ is then given by

$$\begin{aligned}\widehat{\mathbf{x}}_{k|k} &= \boldsymbol{\mu}_{X,k} + \mathbf{K}_k(\mathbf{y}_k - \boldsymbol{\mu}_{Y,k})\\ &= \widehat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k(\mathbf{y}_k - \mathbf{C}_k\,\widehat{\mathbf{x}}_{k|k-1})\,,\end{aligned} \tag{4.47}$$

since, at the present time, our best estimate of $\mathbf{X}_k$ is still $\boldsymbol{\mu}_{X,k} = \widehat{\mathbf{x}}_{k|k-1}$, from which results, by using (4.23), that $\boldsymbol{\mu}_{Y,k} = \mathbf{C}_k\,\boldsymbol{\mu}_{X,k} = \mathbf{C}_k\,\widehat{\mathbf{x}}_{k|k-1}$. In order to calculate $\mathbf{K}_k$, given by $\mathbf{K}_k = \boldsymbol{\Sigma}_{XY}\,\boldsymbol{\Sigma}_{YY}^{-1}$ according to (4.8), we need to evaluate the matrices $\boldsymbol{\Sigma}_{YY}$ and $\boldsymbol{\Sigma}_{XY}$. A calculation similar to the proof of (4.24), applied here for $j = k$, yields successively

$$\begin{aligned}\boldsymbol{\Sigma}_{YY} &= \mathrm{Cov}\left\{\mathbf{Y}_k, \mathbf{Y}_j\right\} = \mathrm{E}\left\{\left[\mathbf{C}_k(\mathbf{X}_k - \widehat{\mathbf{x}}_{k|k-1}) + \mathbf{W}_k\right]\left[\mathbf{C}_k(\mathbf{X}_k - \widehat{\mathbf{x}}_{k|k-1}) + \mathbf{W}_k\right]^{\mathrm{T}}\right\}\\ &= \mathbf{C}_k\,\mathrm{E}\left\{(\mathbf{X}_k - \widehat{\mathbf{x}}_{k|k-1})(\mathbf{X}_k - \widehat{\mathbf{x}}_{k|k-1})^{\mathrm{T}}\right\}\mathbf{C}_k^{\mathrm{T}} + \mathrm{E}\left\{\mathbf{W}_k\mathbf{W}_k^{\mathrm{T}}\right\}\\ &= \mathbf{C}_k\,\boldsymbol{\Sigma}_{k|k-1}\,\mathbf{C}_k^{\mathrm{T}} + \overline{\mathbf{R}}_k\end{aligned}$$

where use has been made of (4.44). Similarly:

$$\begin{aligned}\boldsymbol{\Sigma}_{XY} &= \mathrm{Cov}\left\{\mathbf{X}_k, \mathbf{Y}_k\right\} = \mathrm{E}\left\{(\mathbf{X}_k - \widehat{\mathbf{x}}_{k|k-1})(\mathbf{C}_k\mathbf{X}_k - \mathbf{C}_k\,\widehat{\mathbf{x}}_{k|k-1} + \mathbf{W}_k)^{\mathrm{T}}\right\}\\ &= \boldsymbol{\Sigma}_{k|k-1}\,\mathbf{C}_k^{\mathrm{T}}\end{aligned} \tag{4.48}$$

Finally,

$$\mathbf{K}_k = \boldsymbol{\Sigma}_{k|k-1}\,\mathbf{C}_k^{\mathrm{T}}(\mathbf{C}_k\,\boldsymbol{\Sigma}_{k|k-1}\,\mathbf{C}_k^{\mathrm{T}} + \overline{\mathbf{R}}_k)^{-1}\,. \tag{4.49}$$

Now that $\mathbf{K}_k$ is calculated and $\widehat{\mathbf{x}}_{k|k}$ determined by (4.47). It is clear that the variance of the estimation error $\widetilde{\mathbf{X}}_{k|k} = \mathbf{X}_k - \widehat{\mathbf{x}}_{k|k}$, which corresponds to this new estimate, will also have a new value, which will result from the application of (4.10):

$$\mathbf{\Sigma}_{k|k} = \mathrm{Var}\left\{\widetilde{\mathbf{X}}_{k|k}\right\} = \mathbf{\Sigma}_{k|k-1} - \mathbf{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}}(\mathbf{C}_k\,\mathbf{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}} + \overline{\mathbf{R}}_k)^{-1}\mathbf{C}_k\,\mathbf{\Sigma}_{k|k-1}\,, \quad (4.50)$$

where the last term, $\mathbf{\Sigma}_{YX} = \mathbf{C}_k\,\mathbf{\Sigma}_{k|k-1}$, has been obtained by transposition of (4.48) and application of the symmetry property of any covariance matrix.

Remark 4.7. Recall that $\mathbf{\Sigma}_{k|k-1}$ represents the estimation error variance *before* the measurement, and $\mathbf{\Sigma}_{k|k}$ the same variance *after* the measurement. Equation (4.50) shows then that the latter is never superior to the error variance *before* measurement. The term $\mathbf{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}}(\mathbf{C}_k\,\mathbf{\Sigma}_{k\,|\,k-1}\mathbf{C}_k^{\mathrm{T}} + \overline{\mathbf{R}}_k)^{-1}\mathbf{C}_k\,\mathbf{\Sigma}_{k|k-1}$ expresses indeed the diminution of this variance due to the improvement that the measurement $\mathbf{y}_k$ brings to the knowledge of the state $\mathbf{x}_k$.

Remark 4.8. Equivalent expression of $\mathbf{K}_k$:

$$\mathbf{K}_k = \mathbf{\Sigma}_{k|k}\,\mathbf{C}_k^{\mathrm{T}}\,\overline{\mathbf{R}}_k^{-1}\,. \quad (4.51)$$

Proof. Let us right multiply both sides of (4.50) by $\mathbf{C}_k^{\mathrm{T}}\,\overline{\mathbf{R}}_k^{-1}$, and then factor in the right side the term $\mathbf{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}}(\mathbf{C}_k\,\mathbf{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}} + \overline{\mathbf{R}}_k)^{-1}$ to the left and the term $\overline{\mathbf{R}}_k^{-1}$ to the right. This yields successively

$$\begin{aligned}
&\mathbf{\Sigma}_{k|k}\,\mathbf{C}_k^{\mathrm{T}}\overline{\mathbf{R}}_k^{-1}\\
&\quad= \mathbf{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}}\overline{\mathbf{R}}_k^{-1} - \mathbf{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}}(\mathbf{C}_k\,\mathbf{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}} + \overline{\mathbf{R}}_k)^{-1}\mathbf{C}_k\,\mathbf{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}}\overline{\mathbf{R}}_k^{-1}\\
&\quad= \mathbf{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}}(\mathbf{C}_k\,\mathbf{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}} + \overline{\mathbf{R}}_k)^{-1}\left[(\mathbf{C}_k\mathbf{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}} + \overline{\mathbf{R}}_k) - \mathbf{C}_k\,\mathbf{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}}\right]\overline{\mathbf{R}}_k^{-1}\\
&\quad= \mathbf{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}}(\mathbf{C}_k\,\mathbf{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}} + \overline{\mathbf{R}}_k)^{-1}\left[\overline{\mathbf{R}}_k\right]\overline{\mathbf{R}}_k^{-1}
\end{aligned}$$

The last expression is equal to $\mathbf{K}_k$ according to (4.49).

3rd phase: recursive algorithm and improvement of phase 1. In the hypothesis of a recursive algorithm, a possible improvement of the previous equations seems evident. Nothing indeed kept us at step $k-1$, thus in the first phase of this algorithm, from using the *estimation* $\widehat{\mathbf{x}}_{k-1|k-1}$ of the state of step $k-1$, instead of its *prediction* $\overline{\mathbf{x}}_{k-1} = \widehat{\mathbf{x}}_{k-1|k-2}$, to predict the state of step *k*, since the measurement $\mathbf{y}_{k-1}$ had then already occurred. Equation (4.45) is then replaced by:

$$\widehat{\mathbf{x}}_{k|k-1} = \mathbf{\Phi}_{k-1}\,\widehat{\mathbf{x}}_{k-1|k-1}\,. \quad (4.52)$$

A similar reasoning leads to the replacement of (4.46), which describes the propagation of the estimation error variance from step $k-1$ to step k, by the following:

$$\mathbf{\Sigma}_{k|k-1} = \mathbf{\Phi}_{k-1}\, \mathbf{\Sigma}_{k-1|k-1}\, \mathbf{\Phi}_{k-1}^{\mathrm{T}} + \mathbf{E}_{k-1}\, \bar{\mathbf{Q}}_{k-1}\, \mathbf{E}_{k-1}^{\mathrm{T}}, \tag{4.53}$$

since here again we can take benefit from the improvement which occurred in this variance after determination of the new estimate of step $k-1$.

4.3.1.3 Recapitulation: Discrete Kalman Filter

The Kalman filter is thus represented by a set of recurrent equations, which can be grouped in an estimation phase, corresponding to equations (4.47), (4.49) and (4.50), and a prediction phase, which corresponds to equations (4.52) and (4.53), these two last equations being rewritten here for the step $k+1$:

Estimation (or update) phase:

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k(\mathbf{y}_k - \mathbf{C}_k\, \hat{\mathbf{x}}_{k|k-1}) \tag{4.54}$$

$$\mathbf{K}_k = \mathbf{\Sigma}_{k|k-1}\, \mathbf{C}_k^{\mathrm{T}}(\mathbf{C}_k\, \mathbf{\Sigma}_{k|k-1}\, \mathbf{C}_k^{\mathrm{T}} + \bar{\mathbf{R}}_k)^{-1} = \mathbf{\Sigma}_{k|k}\, \mathbf{C}_k^{\mathrm{T}}\, \bar{\mathbf{R}}_k^{-1} \tag{4.55}$$

$$\mathrm{Var}\left\{\tilde{\mathbf{X}}_{k|k}\right\} = \mathbf{\Sigma}_{k|k} = \mathbf{\Sigma}_{k|k-1} - \mathbf{\Sigma}_{k|k-1}\, \mathbf{C}_k^{\mathrm{T}}(\mathbf{C}_k\, \mathbf{\Sigma}_{k|k-1}\, \mathbf{C}_k^{\mathrm{T}} + \bar{\mathbf{R}}_k)^{-1}\mathbf{C}_k\, \mathbf{\Sigma}_{k|k-1} \tag{4.56}$$

$$= (\mathbf{I} - \mathbf{K}_k\mathbf{C}_k)\mathbf{\Sigma}_{k|k-1} \tag{4.57}$$

Prediction (or propagation) phase:

$$\hat{\mathbf{x}}_{k+1|k} = \mathbf{\Phi}_k\, \hat{\mathbf{x}}_{k|k}\,, \quad \hat{\mathbf{x}}_{0|-1} \text{ given} \tag{4.58}$$

$$\mathbf{\Sigma}_{k+1|k} = \mathbf{\Phi}_k\, \mathbf{\Sigma}_{k|k}\, \mathbf{\Phi}_k^{\mathrm{T}} + \mathbf{E}_k\, \bar{\mathbf{Q}}_k\, \mathbf{E}_k^{\mathrm{T}}, \quad \mathbf{\Sigma}_{0|-1} \text{ given} \tag{4.59}$$

4.3.1.4 Definition of Innovation

The estimation at step k of $\mathbf{X}_k$ according to (4.54) takes into account all past measurements and the present measurement $\mathbf{y}_k$. This last measurement brings redundant information; in reality, the new information about the state $\mathbf{X}_k$ is contained in

$$\boldsymbol{\nu}_k = \mathbf{y}_k - \mathbf{C}_k \widehat{\mathbf{x}}_{k|k-1} ,$$

quantity called *innovation*. In other words, the innovation represents the difference between the *present* measurement $\mathbf{y}_k$ and the *predicted* measurement $\mathbf{C}_k \widehat{\mathbf{x}}_{k|k-1}$. Note of course that $\boldsymbol{\nu}_k$ is a realization of the following random variable $\mathbf{N}_k$:

$$\mathbf{N}_k = \mathbf{Y}_k - \mathbf{C}_k \widehat{\mathbf{x}}_{k|k-1} . \tag{4.60}$$

4.3.1.5 Physical Interpretation of the Kalman Filter

The Kalman filter (equations (4.54) and (4.58)) is thus a model of the system, as described by (4.43), with a correction term proportional to the innovation.

It is interesting to establish a parallel with the physical interpretation which has already been given about the Luenberger observer, in Chap. 2. In fact, the Kalman filter is nothing else than a Luenberger observer considered in a stochastic context, thus with addition of noises.

Furthermore, one notes that the proportionality matrix $\mathbf{K}_k$ given by (4.55), also called *gain matrix* of the filter, is essentially the ratio between

- the uncertainty in the state, $\boldsymbol{\Sigma}_{k|k}$,
- and the uncertainty in the measurement, $\overline{\mathbf{R}}_k$,

the matrix $\mathbf{C}_k$ expressing simply the sensors which build the measurement $\mathbf{Y}_k$ from the state $\mathbf{X}_k$.

This permits understanding better the operation of the Kalman filter, underlying to (4.54):

- if the measurement noise is important, the gain $\mathbf{K}_k$ will become very small, and priority will be given to the model simulation;
- conversely, if the state equation disturbances are important, a gain $\mathbf{K}_k$ which becomes then high will strengthen the influence of the measurements on the estimations.

4.3.1.6 Condensed Form of the Estimation and Prediction Equations

There are two ways of condensing the two previous phases, of estimation and prediction, in one unique phase.

Estimator filter. This form is obtained by elimination of $\widehat{\mathbf{x}}_{k|k-1}$ between (4.54) and (4.58), which yields:

$$\widehat{\mathbf{x}}_{k|k} = \mathbf{\Phi}_{k-1}\,\widehat{\mathbf{x}}_{k-1|k-1} + \mathbf{K}_k(\mathbf{y}_k - \mathbf{C}_k\mathbf{\Phi}_{k-1}\,\widehat{\mathbf{x}}_{k-1|k-1}) \ . \tag{4.61}$$

Joined to the equations (4.55), (4.57) and (4.59), this form leads to the symbolic block diagram of the discrete Kalman filter, illustrated in Fig. 4.5.

Predictor filter. This is the most interesting condensed form for practical applications. By eliminating this time $\widehat{\mathbf{x}}_{k|k}$ between (4.54) and (4.58), we obtain:

$$\widehat{\mathbf{x}}_{k+1|k} = \mathbf{\Phi}_k\,\widehat{\mathbf{x}}_{k|k-1} + \mathbf{\Phi}_k\,\mathbf{K}_k(\mathbf{y}_k - \mathbf{C}_k\widehat{\mathbf{x}}_{k|k-1}) \ . \tag{4.62}$$

Likewise, the elimination $\mathbf{\Sigma}_{k|k}$ between (4.56) and (4.59) yields:

$$\mathbf{\Sigma}_{k+1|k} = \mathbf{\Phi}_k\mathbf{\Sigma}_{k|k-1}\mathbf{\Phi}_k^{\mathrm{T}} - \mathbf{\Phi}_k\mathbf{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}}(\mathbf{C}_k\,\mathbf{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}} + \overline{\mathbf{R}}_k)^{-1}\mathbf{C}_k\,\mathbf{\Sigma}_{k|k-1}\mathbf{\Phi}_k^{\mathrm{T}} \\ + \mathbf{E}_k\overline{\mathbf{Q}}_k\mathbf{E}_k^{\mathrm{T}} \ .(4.63)$$

This equation is again a *Riccati (matrix) difference equation*, very similar to the one already encountered in Chap. 3. The duality between these two equations will be discussed again in Sect. 4.6. This duality allows showing that its solution $\mathbf{\Sigma}_{k+1|k}$ is a positive semidefinite matrix. A criterion of positive semidefiniteness is given in Appendix B**.**

4.3.1.7 Filter Implementation

Initialization. Consider that the measurements are received from the time $t = 0$. Suppose further given the expectation and variance of the random initial state $\mathbf{X}_0$:

$$\mathrm{E}\{\mathbf{X}_0\} = \boldsymbol{\mu}_{X,0} \ ; \quad \mathrm{Var}\{\mathbf{X}_0\} = \mathrm{E}\left\{(\mathbf{X}_0 - \boldsymbol{\mu}_{X,0})(\mathbf{X}_0 - \boldsymbol{\mu}_{X,0})^{\mathrm{T}}\right\} = \mathbf{\Sigma}_0 \, .$$

Two cases can then occur:

1. If the first available measurement is $\mathbf{y}_0$, choose $\widehat{\mathbf{x}}_{0|-1} = \boldsymbol{\mu}_{X,0}$ and $\mathbf{\Sigma}_{0|-1} = \mathbf{\Sigma}_0$; the first step will then be that of estimation at $k = 0$ by using successively (4.55), (4.54) and (4.57); this is the situation depicted in Fig. 4.5;

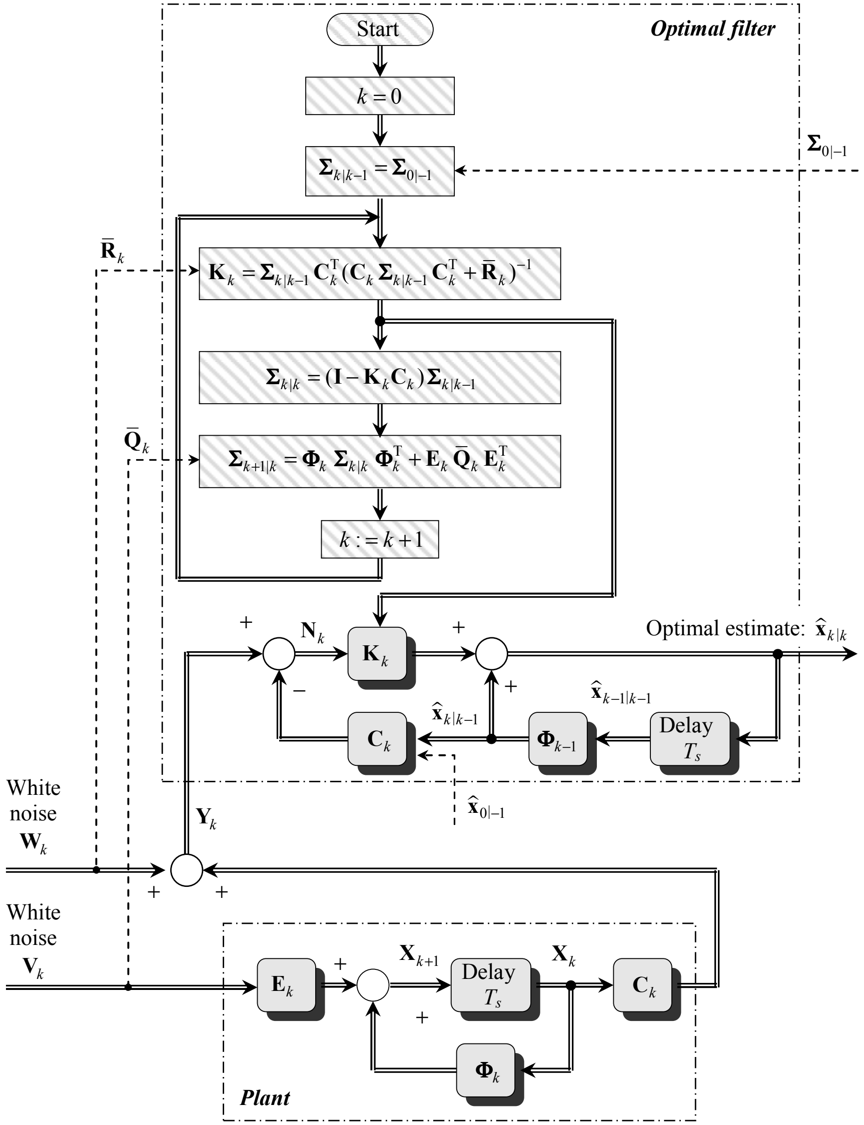

Optimal filter
Start
$k = 0$
$\boldsymbol{\Sigma}_{k|k-1} = \boldsymbol{\Sigma}_{0|-1}$
$\boldsymbol{\Sigma}_{0|-1}$
$\bar{\mathbf{R}}_k$
$\mathbf{K}_k = \boldsymbol{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}}(\mathbf{C}_k\boldsymbol{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}} + \bar{\mathbf{R}}_k)^{-1}$
$\boldsymbol{\Sigma}_{k|k} = (\mathbf{I} - \mathbf{K}_k\mathbf{C}_k)\boldsymbol{\Sigma}_{k|k-1}$
$\bar{\mathbf{Q}}_k$
$\boldsymbol{\Sigma}_{k+1|k} = \boldsymbol{\Phi}_k\boldsymbol{\Sigma}_{k|k}\boldsymbol{\Phi}_k^{\mathrm{T}} + \mathbf{E}_k\bar{\mathbf{Q}}_k\mathbf{E}_k^{\mathrm{T}}$
$k := k + 1$
+
$\mathbf{N}_k$
$\mathbf{K}_k$
+
Optimal estimate: $\hat{\mathbf{x}}_{k|k}$
−
+
$\hat{\mathbf{x}}_{k-1|k-1}$
$\hat{\mathbf{x}}_{k|k-1}$
$\mathbf{C}_k$
$\boldsymbol{\Phi}_{k-1}$
Delay T_s
$\hat{\mathbf{x}}_{0|-1}$
White noise $\mathbf{W}_k$
$\mathbf{Y}_k$
+
+
White noise $\mathbf{V}_k$
$\mathbf{E}_k$
+
$\mathbf{X}_{k+1}$
Delay T_s
$\mathbf{X}_k$
$\mathbf{C}_k$
+
$\boldsymbol{\Phi}_k$
Plant

Fig. 4.5 Functional and algorithmic diagram of a discrete Kalman filter, associated with a discrete plant with disturbances.

2. If the first available measurement is $\mathbf{y}_1$, choose $\hat{\mathbf{x}}_{0|0} = \boldsymbol{\mu}_{X,0}$ and $\boldsymbol{\Sigma}_{0|0} = \boldsymbol{\Sigma}_0$; the first equations used will then be those of prediction of step $k = 1$: equations (4.58) and (4.59) applied for $k = 0$.

Preliminary computations. Neither $\boldsymbol{\Sigma}_{k|k}$ nor $\mathbf{K}_k$ depend on the measurement $\mathbf{y}_k$. They can therefore be calculated in advance and stored in the computer memory, in the form of sequences $\left\{\mathbf{\Sigma}_{k|k}\right\}$ and $\left\{\mathbf{K}_k\right\}$. The computation of $\widehat{\mathbf{x}}_{k|k}$ from $\mathbf{y}_k$ is then very simple, which facilitates the real time implementation of the filter.

4.3.1.8 Properties of the Innovation Sequence

The sequence of innovations $\left\{\mathbf{N}_1,\mathbf{N}_2,\ldots,\mathbf{N}_k\right\}$, generated by the Kalman filter, constitutes a discrete *white noise*, of variance

$$\operatorname{Var}\left\{\mathbf{N}_k\right\}=\mathbf{C}_k\mathbf{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}}+\overline{\mathbf{R}}_k\,. \tag{4.64}$$

Proof. The first result stems from the fact that $\mathbf{N}_k$ depends only on $\mathbf{Y}_k$, and not on the past measurements $\left\{\mathbf{Y}_0,\mathbf{Y}_1,\ldots,\mathbf{Y}_{k-1}\right\}$. On the other hand, $\mathbf{N}_k$ is also independent of the quantities $\left\{\widehat{\mathbf{x}}_{1|0},\widehat{\mathbf{x}}_{2|1},\ldots,\widehat{\mathbf{x}}_{k|k-1}\right\}$, which are functions of the above measurements through (4.62). $\mathbf{N}_k$ is therefore independent of the past innovations $\mathbf{N}_j=\mathbf{Y}_j-\mathbf{C}_j\widehat{\mathbf{x}}_{j|j-1}$, $\forall\, j<k$. In conclusion, $\mathbf{N}_k$ is a white sequence:

$$\operatorname{Cov}\left\{\mathbf{N}_k,\mathbf{N}_j\right\}=0,\quad \forall\, j\neq k\,.$$

This result is intuitive, since if $\mathbf{N}_{k+1}$ were correlated with $\mathbf{N}_k$, this would mean that not all the information available to elaborate $\widehat{\mathbf{x}}_{k|k}$ would have been extracted from $\mathbf{y}_k$, and that this estimate was therefore not optimal. A similar reasoning has already been made to give a physical interpretation of the principle of orthogonality, in Sect. 4.1.3.

It is thus possible in theory to check the optimality of a filter by applying an experimental test of whiteness to the observed innovation.

Furthermore, $\mathbf{N}_k$ can also be rewritten as $\mathbf{N}_k=\mathbf{C}_k(\mathbf{X}_k-\widehat{\mathbf{x}}_{k|k-1})+\mathbf{W}_k$, which proves, with the help of (4.19) and of (4.24), the second half of the stated property.

4.3.2 Continuous Kalman Filter

4.3.2.1 Position of the Problem

Consider a continuous-time linear system, time-varying in the most general case, and submitted to process noise as well as to measurement noise.

It is represented by the following equations:

$$\begin{cases}\dot{\mathbf{X}}(t)=\mathbf{A}(t)\mathbf{X}(t)+\mathbf{E}(t)\mathbf{V}(t)\\ \mathbf{Y}(t)=\mathbf{C}(t)\mathbf{X}(t)+\mathbf{W}(t)\end{cases} \tag{4.65}$$

where again $\mathbf{V}(t)$ and $\mathbf{W}(t)$ are supposed to be white noises, here continuous-time, not correlated with each other. In addition, $\mathbf{V}(t)$ is supposed not correlated with past values of $\mathbf{X}(t)$ and $\mathbf{W}(t)$ not correlated with $\mathbf{X}(t)$ at any time.

The problem hypotheses are thus:

$$\left.\begin{array}{l}\mathrm{E}\{\mathbf{X}_0\}=\boldsymbol{\mu}_{X,0}\,,\quad \mathrm{Var}\{\mathbf{X}_0\}=\boldsymbol{\Sigma}_0\\ \mathrm{E}\{\mathbf{V}(t)\}=\mathrm{E}\{\mathbf{W}(t)\}=0\\ \mathrm{E}\left\{\begin{pmatrix}\mathbf{V}(t)\\ \mathbf{W}(t)\end{pmatrix}\begin{pmatrix}\mathbf{V}(t')\\ \mathbf{W}(t')\end{pmatrix}^{\mathrm{T}}\right\}=\begin{pmatrix}\bar{\mathbf{Q}}(t) & \mathbf{0}\\ \mathbf{0} & \bar{\mathbf{R}}(t)\end{pmatrix}\delta(t-t')\\ \mathrm{E}\left\{\mathbf{V}(t)\mathbf{X}^{\mathrm{T}}(t')\right\}=0,\quad \forall\, t'<t\\ \mathrm{E}\left\{\mathbf{W}(t)\mathbf{X}^{\mathrm{T}}(t')\right\}=0,\quad \forall\, t,t'\end{array}\right\} \tag{4.66}$$

As in the discrete case (Sect. 4.3.1), $\bar{\mathbf{Q}}(t)$ is a symmetric, positive semidefinite matrix (see Sect. C.2.5), and we assume furthermore that $\bar{\mathbf{R}}(t)$, also symmetric, is positive definite. Again we assume that the pair $(\mathbf{A}(t),\mathbf{C}(t))$ is detectable $\forall\, t$, i.e. that the model does not contain any unstable *and* unobservable eigenvalue.

Let us determine the best estimate $\widehat{\mathbf{x}}(t)$ of the state $\mathbf{X}(t)$, in the sense of the minimal estimation error variance, from the measurements $\mathbf{y}(\tau), \tau\in[0,t]$.

4.3.2.2 Derivation of the Equations by Passage to the Limit

Kalman gain matrix. Let us recall here (4.55) of the discrete case:

$$\mathbf{K}_k=\boldsymbol{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}}(\mathbf{C}_k\,\boldsymbol{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}}+\bar{\mathbf{R}}_k)^{-1}\,.$$

The variances of the estimation error before and after the measurement $\boldsymbol{\Sigma}_{k|k-1}$ and $\boldsymbol{\Sigma}_{k|k}$ of the discrete case will be replaced here by $\boldsymbol{\Sigma}(t^-)$ and $\boldsymbol{\Sigma}(t^+)$, where t^- and t^+ will denote now instants, respectively just before and just after the measurement $\mathbf{y}(t)$ occurs.

It is clear that, when the time interval dt separating these two instants will tend to zero, t^- and t^+ will merge with each other, and that

$$\boldsymbol{\Sigma}(t^-) \to \boldsymbol{\Sigma}(t^+) \to \boldsymbol{\Sigma}(t)$$

will hold, i.e. also:

$$\boldsymbol{\Sigma}_{k|k-1} \to \boldsymbol{\Sigma}_{k|k} \to \boldsymbol{\Sigma}(t)\,.$$

By passage to the limit, the following will hold for $\mathbf{K}_k$, by using (4.33) for $\overline{\mathbf{R}}_k$:

$$\begin{aligned}\mathbf{K}_k &\longrightarrow \boldsymbol{\Sigma}(t)\mathbf{C}^{\mathrm{T}}(t)\left[\mathbf{C}(t)\boldsymbol{\Sigma}(t)\mathbf{C}^{\mathrm{T}}(t)+\frac{\overline{\mathbf{R}}(t)}{dt}\right]^{-1}\\ &\longrightarrow \boldsymbol{\Sigma}(t)\mathbf{C}^{\mathrm{T}}(t)\left[\mathbf{C}(t)\boldsymbol{\Sigma}(t)\mathbf{C}^{\mathrm{T}}(t)dt+\overline{\mathbf{R}}(t)\right]^{-1} dt\\ &\longrightarrow \boldsymbol{\Sigma}(t)\mathbf{C}^{\mathrm{T}}(t)\overline{\mathbf{R}}^{-1}(t)\,dt\end{aligned}$$

The last line of this passage to the limit was obtained by neglecting the infinitesimally small term in dt in the brackets before the finite term $\overline{\mathbf{R}}(t)$. The following equivalence relation, between discrete and continuous case, is thus obtained for the Kalman gain:

$$\mathbf{K}_k \longrightarrow \mathbf{K}(t)\,dt\ , \tag{4.67}$$

with:

$$\mathbf{K}(t)=\boldsymbol{\Sigma}(t)\,\mathbf{C}^{\mathrm{T}}(t)\,\overline{\mathbf{R}}^{-1}(t)\ . \tag{4.68}$$

State estimation. There is therefore no more reason to distinguish here between a prediction and an estimation phase, these two phases having merged. At the passage to the limit, the correspondences will become

$$\begin{aligned}&\widehat{\mathbf{x}}_{k|k-1} \to \widehat{\mathbf{x}}_{k|k} &&\to \widehat{\mathbf{x}}(t)\\ &\widehat{\mathbf{x}}_{k+1|k} \to \widehat{\mathbf{x}}_{k+1|k+1} &&\to \widehat{\mathbf{x}}(t)+d\widehat{\mathbf{x}}(t)\,.\end{aligned}$$

Starting e.g. from the condensed form (4.62) of the predictor filter of the discrete case,

$$\widehat{\mathbf{x}}_{k+1|k}=\boldsymbol{\Phi}_k\,\widehat{\mathbf{x}}_{k|k-1}+\boldsymbol{\Phi}_k\,\mathbf{K}_k\,(\mathbf{y}_k-\mathbf{C}_k\widehat{\mathbf{x}}_{k|k-1})\,,$$

the following is obtained, by applying the passage relations (4.29) and (4.30):

$$\widehat{\mathbf{x}}(t)+d\,\widehat{\mathbf{x}}(t)=\left[\mathbf{I}+\mathbf{A}(t)\,dt\right]\widehat{\mathbf{x}}(t)+\left[\mathbf{I}+\mathbf{A}(t)\,dt\right]\mathbf{K}(t)\,dt\left[\mathbf{y}(t)-\mathbf{C}(t)\widehat{\mathbf{x}}(t)\right].$$

By subtracting $\widehat{\mathbf{x}}(t)$ to the two sides and neglecting in the right side the terms which are infinitesimally small in dt^2, we get finally:

$$\frac{d\,\widehat{\mathbf{x}}(t)}{dt}=\dot{\widehat{\mathbf{x}}}(t)=\mathbf{A}(t)\widehat{\mathbf{x}}(t)+\mathbf{K}(t)\left[\mathbf{y}(t)-\mathbf{C}(t)\widehat{\mathbf{x}}(t)\right]. \qquad (4.69)$$

Continuous propagation of the covariance. Let us start again from the condensed form (4.63) of the discrete case:

$$\mathbf{\Sigma}_{k+1|k}=\mathbf{\Phi}_k\mathbf{\Sigma}_{k|k-1}\mathbf{\Phi}_k^{\mathrm{T}}-\mathbf{\Phi}_k\mathbf{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}}(\mathbf{C}_k\mathbf{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}}+\bar{\mathbf{R}}_k)^{-1}\mathbf{C}_k\mathbf{\Sigma}_{k|k-1}\mathbf{\Phi}_k^{\mathrm{T}}+\mathbf{E}_k\bar{\mathbf{Q}}_k\mathbf{E}_k^{\mathrm{T}}.$$

As above, since the two phases have merged, the following correspondences can be established at the passage to the limit:

$$\begin{aligned}\mathbf{\Sigma}_{k|k-1}&\to\mathbf{\Sigma}_{k|k} &&\to\mathbf{\Sigma}(t)\\ \mathbf{\Sigma}_{k+1|k}&\to\mathbf{\Sigma}_{k+1|k+1} &&\to\mathbf{\Sigma}(t)+d\mathbf{\Sigma}(t).\end{aligned}$$

By using the passage relations (4.30), (4.31), (4.33) and (4.67), by omitting temporarily the time dependencies to simplify the writings, and by neglecting again the terms which are infinitesimally small in *dt* of order greater than one, the following equations result successively:

$$\begin{aligned}\mathbf{\Sigma}+d\mathbf{\Sigma}&=(\mathbf{I}+\mathbf{A}\,dt)(\mathbf{I}-\mathbf{K}\mathbf{C}\,dt)\mathbf{\Sigma}(\mathbf{I}+\mathbf{A}\,dt)^{\mathrm{T}}+\mathbf{E}\,dt\frac{\bar{\mathbf{Q}}}{dt}\mathbf{E}^{\mathrm{T}}dt\\ &=(\mathbf{I}+\mathbf{A}\,dt-\mathbf{K}\mathbf{C}\,dt)\,\mathbf{\Sigma}\,(\mathbf{I}+\mathbf{A}\,dt)^{\mathrm{T}}+\mathbf{E}\,dt\frac{\bar{\mathbf{Q}}}{dt}\mathbf{E}^{\mathrm{T}}dt\\ &=\mathbf{\Sigma}+\mathbf{A}\mathbf{\Sigma}\,dt+\mathbf{\Sigma}\mathbf{A}^{\mathrm{T}}dt-\mathbf{K}\mathbf{C}\mathbf{\Sigma}\,dt+\mathbf{E}\bar{\mathbf{Q}}\mathbf{E}^{\mathrm{T}}dt\end{aligned}$$

Hence finally, after reintegration of the time dependencies:

$$\dot{\mathbf{\Sigma}}(t)=\mathbf{A}(t)\mathbf{\Sigma}(t)+\mathbf{\Sigma}(t)\mathbf{A}^{\mathrm{T}}(t)-\mathbf{\Sigma}(t)\mathbf{C}^{\mathrm{T}}(t)\bar{\mathbf{R}}^{-1}(t)\mathbf{C}(t)\mathbf{\Sigma}(t)+\mathbf{E}(t)\bar{\mathbf{Q}}(t)\mathbf{E}^{\mathrm{T}}(t)\ .(4.70)$$

This differential equation, which is nonlinear in $\mathbf{\Sigma}$, is a *matrix differential Riccati equation*, also already encountered in Chap. 3. It can be shown here also that its solution $\mathbf{\Sigma}(t)$ is a positive semidefinite matrix.

Some physical interpretation can be given for this differential equation. The time variation of the error variance, i.e. of the estimation uncertainty, is due to the contribution of three terms:

1. $\mathbf{A}(t)\mathbf{\Sigma}(t)+\mathbf{\Sigma}(t)\mathbf{A}^{\mathrm{T}}(t)$ results from the behavior of the homogeneous system (free response) without measurement;
2. $\bar{\mathbf{Q}}(t)$ expresses the uncertainty increase due to the process noise (state noise);
3. $-\mathbf{\Sigma}(t)\mathbf{C}^{\mathrm{T}}(t)\bar{\mathbf{R}}^{-1}(t)\mathbf{C}(t)\mathbf{\Sigma}(t)$ expresses the uncertainty diminution resulting from the measurements. In the absence of measurement, indeed, the following equation is obtained:

$$\dot{\mathbf{\Sigma}}(t)=\mathbf{A}(t)\mathbf{\Sigma}(t)+\mathbf{\Sigma}(t)\mathbf{A}^{\mathrm{T}}(t)+\mathbf{E}(t)\bar{\mathbf{Q}}(t)\mathbf{E}^{\mathrm{T}}(t)\,,$$

which is the evolution equation of the variance (4.37), describing its free evolution, without improvement term.

4.3.2.3 Innovation

Similarly to the discrete case, it can be shown that the innovation

$$\mathbf{N}(t)=\mathbf{Y}(t)-\mathbf{C}(t)\hat{\mathbf{x}}(t)$$

is a white noise, thus centered, and that its variance equals the one of the measurement noise

$$\operatorname{Cov}\left\{\mathbf{N}(t),\mathbf{N}(t')\right\}=\bar{\mathbf{R}}(t)\,\delta(t-t')\,.$$

It can thus be used here also to test the optimality of the filter (see Sect. 4.3.1.8).

4.3.2.4 Recapitulation

The continuous-time optimal filter, also called *Kalman-Bucy* filter, is thus given by the following equations:

Equations:

$$\dot{\hat{\mathbf{x}}}(t)=\mathbf{A}(t)\hat{\mathbf{x}}(t)+\mathbf{K}(t)\left[\mathbf{y}(t)-\mathbf{C}(t)\hat{\mathbf{x}}(t)\right] \tag{4.71}$$

$$\mathbf{K}(t)=\mathbf{\Sigma}(t)\,\mathbf{C}^{\mathrm{T}}(t)\,\bar{\mathbf{R}}^{-1}(t) \tag{4.72}$$

$$\dot{\mathbf{\Sigma}}(t)=\mathbf{A}(t)\mathbf{\Sigma}(t)+\mathbf{\Sigma}(t)\mathbf{A}^{\mathrm{T}}(t)-\mathbf{\Sigma}(t)\mathbf{C}^{\mathrm{T}}(t)\bar{\mathbf{R}}^{-1}(t)\mathbf{C}(t)\mathbf{\Sigma}(t)+\mathbf{E}(t)\bar{\mathbf{Q}}(t)\mathbf{E}^{\mathrm{T}}(t) \tag{4.73}$$

Initial conditions:

$$\widehat{\mathbf{x}}(0) = \boldsymbol{\mu}_{X,0}$$
$$\boldsymbol{\Sigma}(0) = \boldsymbol{\Sigma}_0$$

Remark 4.9: Initialization. If the initial state is perfectly known, the choice should be of course $\boldsymbol{\mu}_{X,0} = \mathbf{x}_0$ and $\boldsymbol{\Sigma}_0 = 0$. If this is not the case, a plausible initial state should be chosen and a very large matrix $\boldsymbol{\Sigma}_0$ should be associated with it.

The representation of such a filter in Fig. 4.6 shows that it realizes a simulation of the plant (without noise), calculates the estimated quantities $\widehat{\mathbf{x}}$ and $\widehat{\mathbf{y}}$, compares $\widehat{\mathbf{y}}$ to the true measured value **y**, and feeds back the difference into the filter through the gain matrix, **K**.

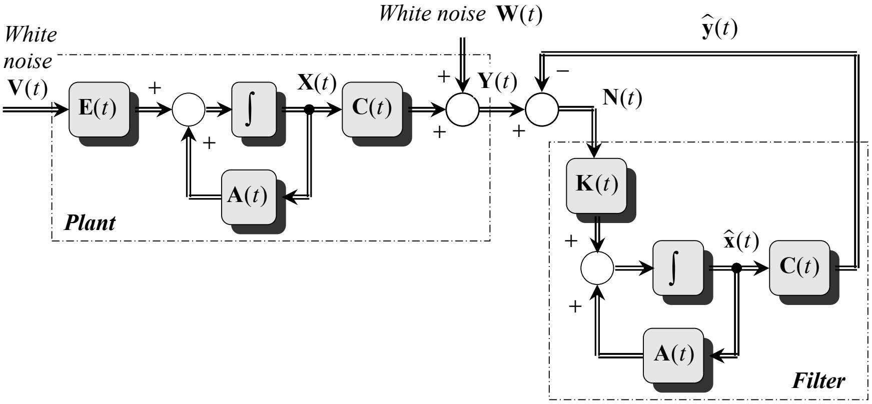


Fig. 4.6 Continuous-time system and associated continuous-time Kalman filter.

4.3.3 Asymptotic Behavior of the Kalman Filter

We will consider now the stationary case, where

- the plant is (stable and) time-invariant: **A** (or $\boldsymbol{\Phi}$), **E**, and **C** are constant;
- the noises are stationary processes: $\overline{\mathbf{Q}}$ and $\overline{\mathbf{R}}$ are constant.

In such a case, the filter can reach, after extinction of the transient phase, a steady state, in the sense that $\boldsymbol{\Sigma}(t)$ becomes a constant matrix, $\dot{\boldsymbol{\Sigma}}(t) = 0$, for the continuous filter, while $\boldsymbol{\Sigma}_{k+1|k} = \boldsymbol{\Sigma}_{k|k-1} = \boldsymbol{\Sigma}^-$ and $\boldsymbol{\Sigma}_{k+1|k+1} = \boldsymbol{\Sigma}_{k|k} = \boldsymbol{\Sigma}^+$ for the discrete filter.

In the continuous case, the differential Riccati equation (4.70) is then replaced by the following *algebraic Riccati equation*:

$$\mathbf{A\Sigma}+\mathbf{\Sigma A}^{\mathrm{T}}-\mathbf{\Sigma C}^{\mathrm{T}}\bar{\mathbf{R}}^{-1}\mathbf{C\Sigma}+\mathbf{E}\bar{\mathbf{Q}}\mathbf{E}^{\mathrm{T}}=0 \ , \tag{4.74}$$

in which all the time dependencies have disappeared.

In the discrete case, the difference equation (4.63) becomes the following *algebraic equation*:

$$\mathbf{\Sigma}^{-}=\mathbf{\Phi\Sigma}^{-}\mathbf{\Phi}^{\mathrm{T}}-\mathbf{\Phi\Sigma}^{-}\mathbf{C}^{\mathrm{T}}(\mathbf{C\Sigma}^{-}\mathbf{C}^{\mathrm{T}}+\bar{\mathbf{R}})^{-1}\mathbf{C\Sigma}^{-}\mathbf{\Phi}^{\mathrm{T}}+\mathbf{E}\bar{\mathbf{Q}}\mathbf{E}^{\mathrm{T}} \ . \tag{4.75}$$

Remark 4.10: Convergence conditions of the solution of these two equations. By duality with the optimal control, the two previous equations have, each respectively, one unique positive semidefinite solution if [Duc02]:

1. the pair $(\mathbf{C},\mathbf{A})$, respectively $(\mathbf{C},\mathbf{\Phi})$, is detectable;
2. the pair $(\mathbf{A},\bar{\mathbf{Q}}_0)$, respectively $(\mathbf{\Phi},\bar{\mathbf{Q}}_0)$, is stabilizable, where $\bar{\mathbf{Q}}_0$ is an arbitrary matrix satisfying $\mathbf{E}\bar{\mathbf{Q}}\mathbf{E}^{\mathrm{T}}=\bar{\mathbf{Q}}_0\bar{\mathbf{Q}}_0^{\mathrm{T}}$.

If in addition $(\mathbf{A},\bar{\mathbf{Q}}_0)$, respectively $(\mathbf{\Phi},\bar{\mathbf{Q}}_0)$, is controllable, $\mathbf{\Sigma}$, respectively $\mathbf{\Sigma}^{-}$, is positive definite.

Remark 4.11. The solution of the two equations permits in theory to calculate $\mathbf{\Sigma}$ or $\mathbf{\Sigma}^{-}$ in steady state. In practice it is however often preferable to integrate numerically (4.70), in the continuous case, or to solve iteratively (4.63) in the discrete case, with $\mathbf{\Sigma}_0=0$, until the steady state solution is reached. Another method consists in using, by duality, the Hamiltonian matrix method discussed in Sect. 3.5.6.

According to (4.72) in the continuous case or (4.55) in the discrete case, the following equations hold in steady state:

$$\left.\begin{array}{r}\mathbf{K}(t)\\ \text{or } \mathbf{K}_k\end{array}\right\}\longrightarrow \mathbf{K}=\begin{cases}\mathbf{\Sigma C}\bar{\mathbf{R}}^{-1} & \text{(continuous)}\\ \mathbf{\Sigma}^{-}\mathbf{C}^{\mathrm{T}}(\mathbf{C\Sigma}^{-}\mathbf{C}^{\mathrm{T}}+\bar{\mathbf{R}})^{-1}=\mathbf{\Sigma}^{+}\mathbf{C}^{\mathrm{T}}\bar{\mathbf{R}}^{-1} & \text{(discrete)}\end{cases} \tag{4.76}$$

where $\mathbf{K}$ is a constant matrix. In the discrete case, the estimation error variance after measurement becomes then:

$$\mathbf{\Sigma}^{+}=(\mathbf{I}-\mathbf{KC})\mathbf{\Sigma}^{-} \ . \tag{4.77}$$

The optimal filter is then time-invariant and is governed, in the continuous case, by the following equation:

$$\dot{\hat{\mathbf{x}}}(t) = (\mathbf{A} - \mathbf{KC})\hat{\mathbf{x}}(t) + \mathbf{K}\,\mathbf{y}(t) \ . \tag{4.78}$$

This equation reflects, in the time domain, the Wiener filter, generated in the frequency domain by solving the Wiener-Hopf integral equation. It is possible to say therefore that the Kalman filter represents an extension of the Wiener filter to time-varying systems and to non stationary noise sources.

Remark 4.12. The constant value $\mathbf{K}$ of the Kalman gain corresponding to the steady state is sometimes used in place of the exact solution $\mathbf{K}_k$ or $\mathbf{K}(t)$, if it is desired to simplify the on-line computations. It should be noted however that such a *stationary* filter will have a different behavior than the Kalman filter described previously, during the transient phase resulting from the initialization or from any sudden change occurring later in the plant state or parameters. During such transients, the stationary filter will not weight correctly the information brought by the initial estimate and the first measurements.

4.3.4 Generalization. Addition of a Deterministic Term

The linear systems considered up to now, continuous or discrete, were fed by only one input signal, which was supposed random.

Equally, the direct feedthrough matrix **D**, which adds to the output equation of a state space represented system a term directly proportional to the input, was assumed to vanish.

The following equations show briefly the modifications which take place in the previous equations, when a deterministic input vector $\mathbf{u}_k$ is added to a discrete system,

$$\begin{cases} \mathbf{X}_{k+1} = \mathbf{\Phi}_k\,\mathbf{X}_k + \mathbf{\Gamma}_k\mathbf{u}_k + \mathbf{E}_k\mathbf{V}_k \\ \quad \mathbf{Y}_k = \mathbf{C}_k\,\mathbf{X}_k + \mathbf{D}_k\mathbf{u}_k + \mathbf{W}_k \end{cases} \tag{4.79}$$

and when a deterministic input $\mathbf{u}(t)$ is added to a continuous system:

$$\begin{cases} \dot{\mathbf{X}}(t) = \mathbf{A}(t)\mathbf{X}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{E}(t)\mathbf{V}(t) \\ \mathbf{Y}(t) = \mathbf{C}(t)\mathbf{X}(t) + \mathbf{D}(t)\mathbf{u}(t) + \mathbf{W}(t) \end{cases} \tag{4.80}$$

It is easy to see that the only changes in the equations of the Kalman filter consist in the introduction of a term containing **u** in (4.54) and (4.58) in the discrete case, and in (4.71) in the continuous case. Fig. 4.7 illustrates this situation.

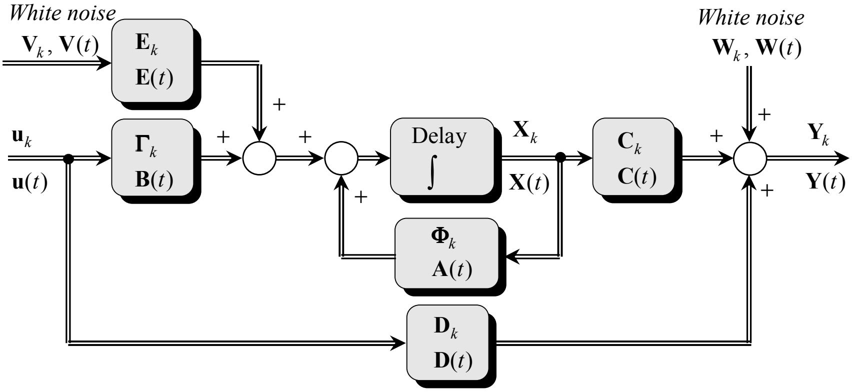


Fig. 4.7 Continuous-time and discrete-time plant with process and measurement white noises and deterministic input.

Let us consider successively the two cases.

Discrete case. The estimation equation (4.54) must be completed as follows:

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k(\mathbf{y}_k - \mathbf{C}_k \hat{\mathbf{x}}_{k|k-1} - \mathbf{D}_k \mathbf{u}_k) \ . \tag{4.81}$$

The corresponding equation yielding the best estimate of the output is:

$$\hat{\mathbf{y}}_{k|k} = \mathbf{C}_k \hat{\mathbf{x}}_{k|k} + \mathbf{D}_k \mathbf{u}_k \ . \tag{4.82}$$

The prediction equation (4.58) becomes

$$\hat{\mathbf{x}}_{k+1|k} = \mathbf{\Phi}_k \hat{\mathbf{x}}_{k|k} + \mathbf{\Gamma}_k \mathbf{u}_k, \quad \hat{\mathbf{x}}_{0|-1} \text{ given} \ , \tag{4.83}$$

and the condensed equation of the *estimator* filter, (4.61), is replaced by

$$\hat{\mathbf{x}}_{k|k} = \mathbf{\Phi}_{k-1} \hat{\mathbf{x}}_{k-1|k-1} + \mathbf{\Gamma}_{k-1}\mathbf{u}_{k-1} + \mathbf{K}_k \Big[\mathbf{y}_k - \mathbf{C}_k(\mathbf{\Phi}_{k-1} \hat{\mathbf{x}}_{k-1|k-1} + \mathbf{\Gamma}_{k-1}\mathbf{u}_{k-1}) - \mathbf{D}_k \mathbf{u}_k\Big] \ , \tag{4.84}$$

whereas that of the *predictor* filter, (4.62), becomes

$$\hat{\mathbf{x}}_{k+1|k} = \mathbf{\Phi}_k \hat{\mathbf{x}}_{k|k-1} + \mathbf{\Gamma}_k \mathbf{u}_k + \mathbf{\Phi}_k \mathbf{K}_k(\mathbf{y}_k - \mathbf{C}_k \hat{\mathbf{x}}_{k|k-1} - \mathbf{D}_k \mathbf{u}_k) \ . \tag{4.85}$$

Continuous case. Equation (4.71) becomes here

$$\dot{\hat{\mathbf{x}}}(t) = \mathbf{A}(t)\hat{\mathbf{x}}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{K}(t)\Big[\mathbf{y}(t) - \mathbf{C}(t)\hat{\mathbf{x}}(t) - \mathbf{D}(t)\mathbf{u}(t)\Big] \;. \tag{4.86}$$

Steady state in the stationary case. Equation (4.78) is replaced by

$$\dot{\hat{\mathbf{x}}}(t) = (\mathbf{A} - \mathbf{KC})\hat{\mathbf{x}}(t) + \mathbf{B}\mathbf{u}(t) + \mathbf{K}\mathbf{y}(t) - \mathbf{K}\mathbf{D}\mathbf{u}(t) \;. \tag{4.87}$$

4.3.5 State Representation of the Discrete Predictor Kalman Filter

The substitution of (4.81) into (4.82) yields:

$$\hat{\mathbf{y}}_{k|k} = \mathbf{C}_k\hat{\mathbf{x}}_{k|k-1} + \mathbf{C}_k\mathbf{K}_k(\mathbf{y}_k - \mathbf{C}_k\hat{\mathbf{x}}_{k|k-1} - \mathbf{D}_k\mathbf{u}_k) + \mathbf{D}_k\mathbf{u}_k \;.$$

Rearranging somewhat the terms of this equation as well as those of (4.85) leads to the following set of two equations:

$$\begin{cases} \hat{\mathbf{x}}_{k+1|k} = \mathbf{\Phi}_k(\mathbf{I} - \mathbf{K}_k\mathbf{C}_k)\hat{\mathbf{x}}_{k|k-1} + (\mathbf{\Gamma}_k - \mathbf{\Phi}_k\mathbf{K}_k\mathbf{D}_k)\mathbf{u}_k + \mathbf{\Phi}_k\mathbf{K}_k\mathbf{y}_k \\ \hat{\mathbf{y}}_{k|k} = \mathbf{C}_k(\mathbf{I} - \mathbf{K}_k\mathbf{C}_k)\hat{\mathbf{x}}_{k|k-1} + (\mathbf{D}_k - \mathbf{C}_k\mathbf{K}_k\mathbf{D}_k)\mathbf{u}_k + \mathbf{C}_k\mathbf{K}_k\mathbf{y}_k \end{cases} \quad (4.88)\begin{cases}(a)\\(b)\end{cases}$$

It is easy to recognize that these two equations are respectively the state equation and the output equation of a linear system, whose state would be the prediction $\hat{\mathbf{x}}_{k|k-1}$ of the state of the observed system and whose output would be the estimate $\hat{\mathbf{y}}_{k|k}$ of the output of this system, after update (estimate after measurement).

The equation yielding the updated estimate of the state of the observed system (equation (4.81) above) can be added to this representation, also after rearrangement:

$$\hat{\mathbf{x}}_{k|k} = (\mathbf{I} - \mathbf{K}_k\mathbf{C}_k)\hat{\mathbf{x}}_{k|k-1} - \mathbf{K}_k\mathbf{D}_k\mathbf{u}_k + \mathbf{K}_k\mathbf{y}_k \;. \tag{4.89}$$

4.3.6 Case of Colored Noise

In all the preceding sections we have assumed that the noise sources applied to the plant, $\mathbf{V}_k$ or $\mathbf{V}(t)$, were white noises.

What to do in the case the input noise is a *colored* noise, therefore a *correlated* noise? This name will be given to a noise whose power density spectrum is no longer flat, thus whose autocorrelation function does not vanish for $\tau \neq 0$ or $i \neq j$.

The solution consists then in reverting to the previous problem, by representing the colored noise process by an appropriate linear system, whose input is fed with white noise (see theorem 4.2 or 4.3). The initial state vector is then augmented, by adding to it the components of this supplementary linear system (or filter). The result is an augmented system, with white noise at its input. The Kalman filter is then calculated for this augmented system.

In fact, if the spectrum of the colored noise is constant over a broad enough spectral bandwidth, as compared with the bandwidth of the considered system, or, equivalently, if its autocorrelation function is narrow enough, as compared to the dominant time constants or response time of this system, this colored noise can be represented to a reasonable approximation by white noise.

4.4 Covariance Matrices of the Noises

4.4.1 General Considerations

In the case of white noises, they are characterized by their covariance matrices:

$$\bar{\mathbf{Q}} = \mathrm{Cov}\left\{\mathbf{V}_k, \mathbf{V}_k\right\} = \mathrm{E}\left\{\mathbf{V}_k \mathbf{V}_k^{\mathrm{T}}\right\}, \text{ and } \bar{\mathbf{R}} = \mathrm{Cov}\left\{\mathbf{W}_k, \mathbf{W}_k\right\} = \mathrm{E}\left\{\mathbf{W}_k \mathbf{W}_k^{\mathrm{T}}\right\}.$$

There is usually no information as to eventual correlations between the components of disturbances acting on the various state variables. Therefore diagonal matrices will be selected for $\bar{\mathbf{Q}}$ and $\bar{\mathbf{R}}$, the diagonal elements of which represent the variances of the components of these disturbances.

4.4.2 Choice of the Process Noise Covariance Matrix

In the majority of situations, the discrete-time plant models stem from the sampling of continuous-time plants at some period T_s. The process noise acts on the continuous-time part of the plant. By assuming it to be white noise, it varies considerably during one sampling period. Therefore, its effect at the end of such a period cannot be determined the same way as we had done for equation (4.59): it is necessary on the contrary to integrate it.

By supposing, in order to simplify the writings and without any loss of generality, that the plant is time-invariant, thus that $\mathbf{\Phi}_k = \mathbf{\Phi}$, $\mathbf{\Gamma}_k = \mathbf{\Gamma}$, $\mathbf{E}(t) = \mathbf{E}$, equation (A.22) of Appendix A, which served to establish the state equation of a sampled data system, namely

$$\mathbf{x}_{k+1} = e^{\mathbf{A}T_s}\,\mathbf{x}_k + \int_{kT_s}^{(k+1)T_s} e^{\mathbf{A}\left[(k+1)T_s-\tau\right]}\,\mathbf{B}\,\mathbf{u}(\tau)\,d\tau\ ,$$

permits calculating the conjugated effect of deterministic and stochastic inputs after one period. Starting from (4.80), the evolution of the state over one period is given by

$$\mathbf{X}_{k+1} = \mathbf{\Phi}\mathbf{X}_k + \mathbf{\Gamma}\mathbf{u}_k + \int_0^{T_s} e^{\mathbf{A}(T_s-\tau)}\mathbf{E}\,\mathbf{V}(\tau)\,d\tau\ ,$$

equation which replaces (4.79). With the use of (4.83), the error of the estimation of $\widehat{\mathbf{x}}_{k+1|k}$ before the measurement $\mathbf{y}_{k+1}$, thus in the prediction phase, can be written:

$$\begin{aligned}\widetilde{\mathbf{X}}_{k+1|k} = \mathbf{X}_{k+1} - \widehat{\mathbf{x}}_{k+1|k} &= \mathbf{\Phi}\mathbf{X}_k + \mathbf{\Gamma}\mathbf{u}_k + \int_0^{T_s} e^{\mathbf{A}(T_s-\tau)}\mathbf{E}\,\mathbf{V}(\tau)\,d\tau - \mathbf{\Phi}\widehat{\mathbf{x}}_{k|k} - \mathbf{\Gamma}\mathbf{u}_k \\ &= \mathbf{\Phi}(\mathbf{X}_k - \widehat{\mathbf{x}}_{k|k}) + \int_0^{T_s} e^{\mathbf{A}(T_s-\tau)}\mathbf{E}\,\mathbf{V}(\tau)\,d\tau \qquad .\end{aligned}$$

The variance of this estimation error is given by $\mathbf{\Sigma}_{k+1|k} = \mathrm{E}\left\{\widetilde{\mathbf{X}}_{k+1|k}\widetilde{\mathbf{X}}_{k+1|k}^{\mathrm{T}}\right\}$, the estimate being unbiased, i.e.

$$\begin{aligned}\mathbf{\Sigma}_{k+1|k} = \mathrm{E}\Bigg\{&\left[\mathbf{\Phi}(\mathbf{X}_k - \widehat{\mathbf{x}}_{k|k}) + \int_0^{T_s} e^{\mathbf{A}(T_s-\tau)}\mathbf{E}\,\mathbf{V}(\tau)\,d\tau\right] \\ &\cdot\left[\mathbf{\Phi}(\mathbf{X}_k - \widehat{\mathbf{x}}_{k|k}) + \int_0^{T_s} e^{\mathbf{A}(T_s-\tau)}\mathbf{E}\,\mathbf{V}(\tau)\,d\tau\right]^{\mathrm{T}}\Bigg\} \\ = \mathbf{\Phi}\mathbf{\Sigma}_{k|k}\mathbf{\Phi}^{\mathrm{T}} &+ \int_0^{T_s}\int_0^{T_s} e^{\mathbf{A}(T_s-\tau)}\mathbf{E}\,\mathrm{E}\left\{\mathbf{V}(\tau)\mathbf{V}^{\mathrm{T}}(\tau')\right\}\mathbf{E}^{\mathrm{T}}e^{\mathbf{A}^{\mathrm{T}}(T_s-\tau')}\,d\tau'd\tau\ . \quad (4.90)\end{aligned}$$

The noise $\mathbf{V}(t)$ being white noise, assumed stationary, it satisfies

$$\mathrm{E}\left\{\mathbf{V}(\tau)\mathbf{V}^{\mathrm{T}}(\tau')\right\} = \overline{\mathbf{Q}}\,\delta(\tau-\tau')\,,$$

where $\overline{\mathbf{Q}}$ is its power spectral density (see Sect. C.2.6). This yields

$$\int_0^{T_s} \mathrm{E}\left\{\mathbf{V}(\tau)\mathbf{V}^{\mathrm{T}}(\tau')\right\}\mathbf{E}^{\mathrm{T}}e^{\mathbf{A}^{\mathrm{T}}(T_s-\tau')}\,d\tau' = \int_0^{T_s} \bar{\mathbf{Q}}\,\delta(\tau-\tau')\mathbf{E}^{\mathrm{T}}e^{\mathbf{A}^{\mathrm{T}}(T_s-\tau')}\,d\tau'$$
$$= \bar{\mathbf{Q}}\mathbf{E}^{\mathrm{T}}e^{\mathbf{A}^{\mathrm{T}}(T_s-\tau)}\,,$$

and (4.90) becomes:

$$\mathbf{\Sigma}_{k+1|k} = \mathbf{\Phi}\mathbf{\Sigma}_{k|k}\mathbf{\Phi}^{\mathrm{T}} + \int_0^{T_s} e^{\mathbf{A}(T_s-\tau)}\mathbf{E}\bar{\mathbf{Q}}\mathbf{E}^{\mathrm{T}}e^{\mathbf{A}^{\mathrm{T}}(T_s-\tau)}d\tau\,.$$

Assuming further that, as is usually the case, the sampling period T_s is significantly shorter than the shortest time constant of the plant, which implies

$$e^{\mathbf{A}(T_s-\tau)} \simeq \mathbf{I}\,,$$

the previous equation becomes:

$$\mathbf{\Sigma}_{k+1|k} \simeq \mathbf{\Phi}\mathbf{\Sigma}_{k|k}\mathbf{\Phi}^{\mathrm{T}} + \mathbf{E}\bar{\mathbf{Q}}\mathbf{E}^{\mathrm{T}}\,T_s\,,$$

or, after reintegration of the eventual time dependencies of the various parameters:

$$\mathbf{\Sigma}_{k+1|k} \simeq \mathbf{\Phi}_k\,\mathbf{\Sigma}_{k|k}\mathbf{\Phi}_k^{\mathrm{T}} + \mathbf{E}(t)\bar{\mathbf{Q}}(t)\mathbf{E}^{\mathrm{T}}(t)\,T_s\,,$$

which, with the equivalence $\mathbf{E}_k = \int_0^{T_s} e^{\mathbf{A}\tau}\mathbf{E}(\tau)d\tau \simeq \mathbf{E}(t)T_s$ for small T_s, can be written

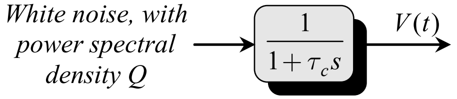
$$\mathbf{\Sigma}_{k+1|k} \simeq \mathbf{\Phi}_k\,\mathbf{\Sigma}_{k|k}\mathbf{\Phi}_k^{\mathrm{T}} + \mathbf{E}_k\,\frac{\bar{\mathbf{Q}}(t)}{T_s}\,\mathbf{E}_k\,. \tag{4.91}$$

By comparison of this equation with (4.59) the passage relation (4.33) has been regained here:

$$\bar{\mathbf{Q}}_k \longrightarrow \frac{\bar{\mathbf{Q}}(t)}{dt}\,.$$

It is then possible to apply the prediction equation of the variance (4.59) of the discrete case to the case where $\mathbf{V}(t)$ is continuous-time noise, by using the approximation $\bar{\mathbf{Q}}_k \simeq \bar{\mathbf{Q}}(t)/T_s$, under the hypothesis that T_s is much smaller than the plant time constants.

The white noise does not exist in reality, where rather “filtered” white noise is encountered. It has been shown that for the system below the following properties are true:



1. Autocorrelation function of the output noise, or real physical noise $V(t)$:

$$\varphi_{VV}(\tau) = \sigma^2 \, e^{-|\tau|/\tau_c} ,$$

where τ_c represents the autocorrelation constant of the physical noise and where

$$\sigma^2 = \varphi_{VV}(0) = \mathrm{E}\left\{V^2(t)\right\} = \mathrm{Var}\left\{V(t)\right\},$$

since $V(t)$ is centered.

2. Relation between the power spectral density of the white noise and the quadratic mean of the physical signal:

$$Q = 2\tau_c\sigma^2 = 2\tau_c \, \mathrm{E}\left\{V^2(t)\right\}.$$

The discrete noise covariance matrix, which enters equation (4.59), will thus have, according to (4.91), the diagonal element *ii*

$$\left(\bar{\mathbf{Q}}_k\right)_{ii} = \frac{2}{T_s}\left[\tau_c \, \mathrm{E}\left\{V^2(t)\right\}\right]_i ,$$

which involves the measured mean quadratic value, and the measured or estimated autocorrelation time constant of the disturbance which acts on the i^{th} component of the state vector.

4.4.3 Choice of the Measurement Noise Covariance Matrix

Here, on the contrary, the noise is most often discrete-time noise, stemming directly from the inaccuracy of the sampled sensors. It is white noise, therefore uncorrelated from sample to sample. The simplest value to take for the diagonal elements of the matrix $\bar{\mathbf{R}}$ is the square of the mean quadratic error of the corresponding sensor, or square of its RMS value,

$$\bar{\mathbf{R}}_{ii} = \left[\mathrm{E}\left\{W^2(t)\right\}\right]_i . \tag{4.92}$$

4.5 State Feedback Including a Kalman Filter (Discrete-time System)

The block diagram of such a control system is shown in Fig. 4.8. The Kalman filter plays the same role as an observer, seen in Chap. 2, with the additional property of filtering noises, in the case of noisy plants.

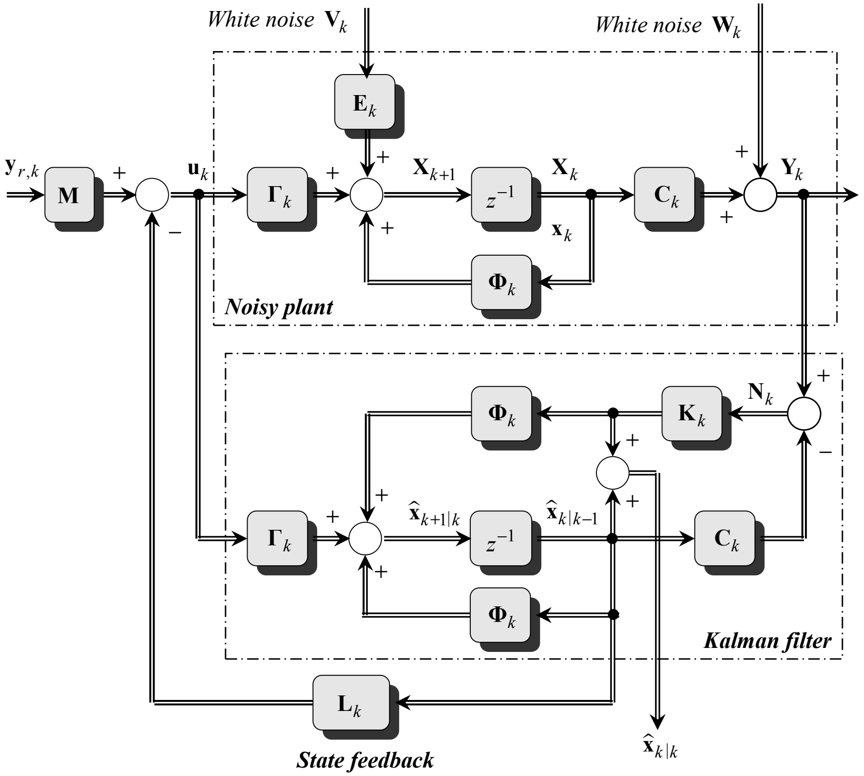


Fig. 4.8 Discrete-time state-feedback control system with Kalman filter in the loop.

4.5.1 Closed-loop Transfer Function

Hypotheses: in this section, we will assume that

1. the plant is linear and time-invariant (constant matrices), with matrix $\mathbf{D} = 0$;
2. the noises vanish;
3. the initial conditions of both the plant and the filter are all zero.

Under these hypotheses, (4.88) yields

$$\widehat{\mathbf{x}}_{k|k-1}=\mathbf{\Phi}(\mathbf{I}-\mathbf{KC})\widehat{\mathbf{x}}_{k-1|k-2}+\mathbf{\Gamma u}_{k-1}+\mathbf{\Phi K y}_{k-1},$$

hence, with $\mathbf{y}_{k-1}=\mathbf{Cx}_{k-1}$,

$$\widehat{\mathbf{x}}_{k|k-1}=\mathbf{\Phi}(\mathbf{I}-\mathbf{KC})\widehat{\mathbf{x}}_{k-1|k-2}+\mathbf{\Gamma u}_{k-1}+\mathbf{\Phi KC x}_{k-1}.$$

In the control law $\mathbf{u}_k=-\mathbf{Lx}_k$, $\mathbf{x}_k$ will be replaced either by $\widehat{\mathbf{x}}_{k|k-1}$ or by $\widehat{\mathbf{x}}_{k|k}$. Let us make the first choice, the most common in practice since it allows computing $\mathbf{u}_k$ *in advance*, during the interval of application of the present control $\mathbf{u}_{k-1}$, from the values available at the present step, $k-1$.

By application of the *z*-transform, with the notations $\mathcal{Z}\left[\widehat{\mathbf{x}}_{k|k-1}\right]=\widehat{\mathbf{X}}^-(z)$, $\mathcal{Z}\left[\mathbf{x}_k\right]=\mathbf{X}(z)$ and $\mathcal{Z}\left[\mathbf{u}_k\right]=\mathbf{U}(z)$, we can write:

$$\widehat{\mathbf{X}}^-(z)=z^{-1}\mathbf{\Phi}(\mathbf{I}-\mathbf{KC})\widehat{\mathbf{X}}^-(z)+z^{-1}\mathbf{\Gamma U}(z)+z^{-1}\mathbf{\Phi KC X}(z). \tag{4.93}$$

On the other hand, by applying the z-transform to (4.79) at $\mathbf{V}_k\equiv 0$,

$$\mathbf{X}(z)=(z\mathbf{I}-\mathbf{\Phi})^{-1}\mathbf{\Gamma U}(z)=z^{-1}(\mathbf{I}-z^{-1}\mathbf{\Phi})^{-1}\mathbf{\Gamma U}(z).$$

Combining these two equations yields

$$\begin{aligned}\left[\mathbf{I}-z^{-1}\mathbf{\Phi}(\mathbf{I}-\mathbf{KC})\right]\widehat{\mathbf{X}}^-(z)&=z^{-1}\mathbf{\Gamma U}(z)+z^{-1}\mathbf{\Phi KC}z^{-1}(\mathbf{I}-z^{-1}\mathbf{\Phi})^{-1}\mathbf{\Gamma U}(z)\\&=z^{-1}\left[\mathbf{I}+\mathbf{\Phi KC}z^{-1}(\mathbf{I}-z^{-1}\mathbf{\Phi})^{-1}\right]\mathbf{\Gamma U}(z)\\&=z^{-1}\left[\mathbf{I}-z^{-1}\mathbf{\Phi}+\mathbf{\Phi KC}z^{-1}\right](\mathbf{I}-z^{-1}\mathbf{\Phi})^{-1}\mathbf{\Gamma U}(z)\\&=z^{-1}\left[\mathbf{I}-z^{-1}\mathbf{\Phi}(\mathbf{I}-\mathbf{KC})\right](\mathbf{I}-z^{-1}\mathbf{\Phi})^{-1}\mathbf{\Gamma U}(z)\quad ,\end{aligned}$$

from where results successively that

$$\widehat{\mathbf{X}}^-(z)=z^{-1}(\mathbf{I}-z^{-1}\mathbf{\Phi})^{-1}\mathbf{\Gamma U}(z)=\mathbf{X}(z), \tag{4.94}$$

$$\widehat{\mathbf{x}}_{k|k-1}=\mathbf{x}_k.$$

The prediction calculated by the Kalman filter “sticks” to the true state. The same should hold for the estimate $\widehat{\mathbf{x}}_{k|k}$. Let us verify it. According to (4.81) or (4.54),

$$\widehat{\mathbf{x}}_{k|k} = \widehat{\mathbf{x}}_{k|k-1} + \mathbf{K}(\mathbf{y}_k - \mathbf{C}\widehat{\mathbf{x}}_{k|k-1}) .$$

Hence, from above:

$$\widehat{\mathbf{x}}_{k|k} = \widehat{\mathbf{x}}_{k|k-1} + \mathbf{K}(\mathbf{C}\mathbf{x}_k - \mathbf{C}\widehat{\mathbf{x}}_{k|k-1}) = \widehat{\mathbf{x}}_{k|k-1} = \mathbf{x}_k .$$

Conclusion. The closed-loop transfer function is the same as if the loop were closed without any Kalman filter included.

This sends back to the discussion in Sect. 2.5.5, concerning the behavior of a closed-loop system including an observer in the loop with respect to reference changes, when the plant and observer initial conditions are identical, or once the estimated state has converged to the true state. This should not constitute a surprise, if one remembers that a Kalman filter is nothing else but an observer for noisy systems (see Sect. 4.3.1.5).

The gain compensation matrix will thus have the same value as calculated without filter. Let us verify this.

4.5.2 Calculation of the Gain Compensation Matrix

By inspection of Fig. 4.8 or by rewriting (4.93) with $\mathbf{u}_k = -\mathbf{L}\widehat{\mathbf{x}}_{k|k-1} + \mathbf{M}\,\mathbf{y}_{r,k}$ and then applying the *z*-transform, we obtain:

$$\widehat{\mathbf{X}^-}(z) = z^{-1}\left[\mathbf{\Phi}(\mathbf{I} - \mathbf{KC}) - \mathbf{\Gamma L}\right]\widehat{\mathbf{X}^-}(z) + z^{-1}\mathbf{\Phi KCX}(z) + z^{-1}\mathbf{\Gamma M Y}_r(z) ,$$

where $\mathcal{Z}\left[\mathbf{y}_{r,k}\right] = \mathbf{Y}_r(z)$. Multiplying now both sides by *z* yields:

$$(z\mathbf{I} - \mathbf{\Phi} + \mathbf{\Phi KC} + \mathbf{\Gamma L})\widehat{\mathbf{X}^-}(z) = \mathbf{\Phi KCX}(z) + \mathbf{\Gamma M Y}_r(z) .$$

Taking into account (4.94), this equation writes also:

$$(z\mathbf{I} - \mathbf{\Phi} + \mathbf{\Phi KC} + \mathbf{\Gamma L})\,\mathbf{X}(z) = \mathbf{\Phi KCX}(z) + \mathbf{\Gamma M Y}_r(z) ,$$

$$(z\mathbf{I} - \mathbf{\Phi} + \mathbf{\Gamma L})\,\mathbf{X}(z) = \mathbf{\Gamma M Y}_r(z) ,$$

hence, finally:

$$\mathbf{Y}(z) = \mathbf{CX}(z) = \mathbf{C}(z\mathbf{I} - \mathbf{\Phi} + \mathbf{\Gamma L})^{-1}\,\mathbf{\Gamma M Y}_r(z) . \tag{4.95}$$

The requirement $\mathbf{Y}(z) = \mathbf{Y}_r(z)$ in response to a constant reference $\mathbf{y}_r$ yields thus, by letting $z = 1$ in (4.95), the following relation for **M**:

$$\mathbf{M} = \left[\mathbf{C}(\mathbf{I} - \mathbf{\Phi} + \mathbf{\Gamma L})^{-1}\,\mathbf{\Gamma}\right]^{-1} .$$

This expression is exactly the same as in the case of a direct state feedback, without filter in the loop, as given by (1.16).

4.6 Principle of Duality

4.6.1 Already Encountered Dualities

Let us recall the very first duality identified in Appendix A, between the controllability and the observability canonical forms of the state equations. The equations of either one of these forms are obtained from the other by simple symmetry of the triplet **A**, **B** and **C** ($\mathbf{\Phi}$**,** $\mathbf{\Gamma}$, and **C**) with respect to the main diagonal of **A** ($\mathbf{\Phi}$):

$$\begin{aligned} \mathbf{A}(\mathbf{\Phi}) &\leftrightarrow \mathbf{A}^{\mathrm{T}}(\mathbf{\Phi}^{\mathrm{T}}) \\ \mathbf{B}(\mathbf{\Gamma}) &\leftrightarrow \mathbf{C}^{\mathrm{T}} \\ \mathbf{C} &\leftrightarrow \mathbf{B}^{\mathrm{T}}(\mathbf{\Gamma}^{\mathrm{T}}) \ . \end{aligned}$$

Later we have remarked that there exists a duality of the same nature between a state feedback and an observer design, thus between a control and an estimation problem.

4.6.2 Duality between Filtering and Control

Now, as mentioned previously, the observer is a particular case of the optimal filtering. It is therefore not surprising to recognize this duality between the optimal linear filtering and the optimal control theories. These two theories have indeed many dual aspects, among which the fact that both require the solving of a matrix Riccati differential, or difference, equation. The essentials of these duality properties are grouped in Table 4.1 for the continuous case, and in Table 4.2 for the discrete case.

In practice, this means that the design of an optimal filter and that of an optimal controller can be made with the same numerical computation programs. This has been used extensively in the creation of the accompanying software, *MMCE.m*.

Table 4.1 Extended duality between filtering and control, for the continuous case.

Optimal filtering	**Optimal control**
$\bar{\mathbf{Q}}$	$\mathbf{Q}^{\mathrm{T}}$
$\bar{\mathbf{R}}$	$\mathbf{R}^{\mathrm{T}}$
$\mathbf{K}$	$\mathbf{L}^{\mathrm{T}}$
$\mathbf{\Sigma}$	$\mathbf{P}^{\mathrm{T}}$
t	$-t$
$\begin{cases} \dot{\mathbf{X}}(t) = \mathbf{A}(t)\mathbf{X}(t) + \mathbf{E}(t)\mathbf{V}(t) \\ \mathbf{Y}(t) = \mathbf{C}(t)\mathbf{X}(t) + \mathbf{W}(t) \end{cases}$	$\begin{cases} \dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) \\ \mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t) \end{cases}$
Optimal estimate:	Optimal state feedback:
$\dot{\hat{\mathbf{x}}}(t) = \mathbf{A}(t)\hat{\mathbf{x}}(t) + \mathbf{K}(t)[\mathbf{y}(t) - \mathbf{C}(t)\hat{\mathbf{x}}(t)]$	$\mathbf{u}(t) = -\mathbf{L}(t)\mathbf{x}(t)$
with:	with:
$\mathbf{K}(t) = \mathbf{\Sigma}(t)\,\mathbf{C}^{\mathrm{T}}(t)\,\bar{\mathbf{R}}^{\mathbf{-1}}(t)\,,$	$\mathbf{L}(t) = \mathbf{R}^{-1}(t)\,\mathbf{B}^{\mathrm{T}}(t)\,\mathbf{P}(t)\,,$
where $\mathbf{\Sigma}(t)$ is solution of the (differential) Riccati equation	where $\mathbf{P}(t)$ is solution of the (differential) Riccati equation
$\begin{aligned} \dot{\mathbf{\Sigma}}(t) = {} & \mathbf{A}(t)\mathbf{\Sigma}(t) + \mathbf{\Sigma}(t)\mathbf{A}^{\mathrm{T}}(t) \\ & -\mathbf{\Sigma}(t)\mathbf{C}^{\mathrm{T}}(t)\bar{\mathbf{R}}^{-1}(t)\mathbf{C}(t)\mathbf{\Sigma}(t) \\ & +\mathbf{E}(t)\bar{\mathbf{Q}}(t)\mathbf{E}^{\mathrm{T}}(t) \end{aligned}$	$\begin{aligned} -\dot{\mathbf{P}}(t) = {} & \mathbf{P}(t)\mathbf{A}(t) + \mathbf{A}^{\mathrm{T}}(t)\mathbf{P}(t) \\ & -\mathbf{P}(t)\mathbf{B}(t)\mathbf{R}^{-1}(t)\mathbf{B}^{\mathrm{T}}(t)\mathbf{P}(t) \\ & +\mathbf{Q}(t) \end{aligned}$

For the discrete case, due to limited space and also to avoid unnecessary duplications, these properties are limited to the Kalman and state-feedback gain matrices, and to the Riccati equations.

Table 4.2 Some duality properties between filtering and control, for the discrete case.

Optimal filtering
$\mathbf{K}_k = \mathbf{\Sigma}_{k\mid k-1}\mathbf{C}_k^{\mathrm{T}}(\mathbf{C}_k\,\mathbf{\Sigma}_{k\mid k-1}\mathbf{C}_k^{\mathrm{T}} + \bar{\mathbf{R}}_k)^{-1} = \mathbf{\Sigma}_{k\mid k}\,\mathbf{C}_k^{\mathrm{T}}\,\bar{\mathbf{R}}_k^{-1}$
$\mathbf{\Sigma}_{k+1\mid k} = \mathbf{\Phi}_k\mathbf{\Sigma}_{k\mid k-1}\mathbf{\Phi}_k^{\mathrm{T}} - \mathbf{\Phi}_k\mathbf{\Sigma}_{k\mid k-1}\mathbf{C}_k^{\mathrm{T}}(\mathbf{C}_k\mathbf{\Sigma}_{k\mid k-1}\mathbf{C}_k^{\mathrm{T}}+\bar{\mathbf{R}}_k)^{-1}\mathbf{C}_k\mathbf{\Sigma}_{k\mid k-1}\mathbf{\Phi}_k^{\mathrm{T}} + \mathbf{E}_k\bar{\mathbf{Q}}_k\mathbf{E}_k^{\mathrm{T}}$
Optimal control
$\mathbf{L}_k = (\mathbf{R}_k + \mathbf{\Gamma}_k^{\mathrm{T}}\mathbf{P}_{k+1}\mathbf{\Gamma}_k)^{-1}\mathbf{\Gamma}_k^{\mathrm{T}}\mathbf{P}_{k+1}\mathbf{\Phi}_k$
$\mathbf{P}_k = \mathbf{\Phi}_k^{\mathrm{T}}\mathbf{P}_{k+1}\,\mathbf{\Phi}_k - \mathbf{\Phi}_k^{\mathrm{T}}\mathbf{P}_{k+1}\,\mathbf{\Gamma}_k(\mathbf{R}_k + \mathbf{\Gamma}_k^{\mathrm{T}}\mathbf{P}_{k+1}\,\mathbf{\Gamma}_k)^{-1}\mathbf{\Gamma}_k^{\mathrm{T}}\mathbf{P}_{k+1}\,\mathbf{\Phi}_k + \mathbf{Q}_k$

4.6.3 Principle of Duality in Mathematics

The following definition gives a broader scope to the above identified duality.

Principle of duality. *Given the following system in state space representation:*

$$\begin{cases}\dot{\mathbf{x}}(t)=\mathbf{A}(t)\mathbf{x}(t)+\mathbf{B}(t)\mathbf{u}(t)\\ \mathbf{y}(t)=\mathbf{C}(t)\mathbf{x}(t)\end{cases} \qquad \mathbf{(I)}$$

the system defined as follows:

$$\begin{cases}-\dot{\boldsymbol{\xi}}(t)=\mathbf{A}^{\mathrm{T}}(t)\boldsymbol{\xi}(t)+\mathbf{C}^{\mathrm{T}}(t)\mathbf{v}(t)\\ \quad\boldsymbol{\eta}(t)=\mathbf{B}^{\mathrm{T}}(t)\boldsymbol{\xi}(t)\end{cases} \qquad \mathbf{(II)}$$

*and which is obtained from the initial system by reversing the time progress, by swapping the input and output matrices and by transposing the **A**, **B** and **C**, is called dual system of the first one.*
*It is worthwhile to note that these two last equations amount to replacing the triplet **A**, **B** and **C** by its symmetrical triplet with respect to the main diagonal of **A**.*
*The estimation and the filtering of the initial system (**I**) correspond to the control of the dual system (**II**), and vice versa.*

Remark 4.13. In differential calculus, the system (**II**) is also called *adjoint system* of system (**I**).

It is easily verified that the application of this principle of duality, due to Kalman (1960), permits retrieving all the *dual* results stated above. By giving the possibility to transpose a number of results from the optimal control theory to the domain of estimation and optimal filtering, he accelerated considerably the development of this second theory, historically more recent.

Let us mention, by curiosity, that the use of the mathematical duality principle yields directly the adjoint equation of a linear system, without resorting to (3.13).

4.6.4 Duality of the Transition Matrices

The previous symmetry rule extends also to the transition matrices of the two systems.

Denote indeed by $\boldsymbol{\Psi}$ the transition matrix of the adjoint system (**II**). It will be solution of the homogeneous differential equation of the corresponding system (see Sect. A.5), thus of

$$\dot{\boldsymbol{\Psi}}(t,t_0)=-\mathbf{A}^{\mathrm{T}}(t)\,\boldsymbol{\Psi}(t,t_0)\,.$$

By grouping all the terms to the left side of this equation and premultiplying by the transpose of the transition matrix $\mathbf{\Phi}(t,t_0)$ of the initial system, the following holds:

$$\mathbf{\Phi}^{\mathrm{T}}(t,t_0)\,\dot{\mathbf{\Psi}}(t,t_0)+\mathbf{\Phi}^{\mathrm{T}}(t,t_0)\mathbf{A}^{\mathrm{T}}(t)\,\mathbf{\Psi}(t,t_0)=0\,. \tag{4.96}$$

Since

$$\dot{\mathbf{\Phi}}(t,t_0)=\mathbf{A}(t)\,\mathbf{\Phi}(t,t_0)\,,$$

it results, by substitution in (4.96), that

$$\mathbf{\Phi}^{\mathrm{T}}(t,t_0)\,\dot{\mathbf{\Psi}}(t,t_0)+\dot{\mathbf{\Phi}}^{\mathrm{T}}(t,t_0)\,\mathbf{\Psi}(t,t_0)=0\,,$$

i.e.

$$\frac{d}{dt}\Big[\mathbf{\Phi}^{\mathrm{T}}(t,t_0)\,\mathbf{\Psi}(t,t_0)\Big]=0\,,$$

or:

$$\mathbf{\Phi}^{\mathrm{T}}(t,t_0)\,\mathbf{\Psi}(t,t_0)=\text{constant}\,.$$

Since we have $\mathbf{\Phi}^{\mathrm{T}}(t_0,t_0)=\mathbf{\Psi}(t_0,t_0)=\mathbf{I}$ at $t=t_0$, this implies that the matrix constant in the right side is also equal to **I**:

$$\mathbf{\Phi}^{\mathrm{T}}(t,t_0)\,\mathbf{\Psi}(t,t_0)=\mathbf{I}\,. \tag{4.97}$$

Consequently,

$$\mathbf{\Psi}(t,t_0)=\Big[\mathbf{\Phi}^{\mathrm{T}}(t,t_0)\Big]^{-1}\,,$$

i.e.:

$$\mathbf{\Psi}(t,t_0)=\mathbf{\Phi}^{\mathrm{T}}(t_0,t) \tag{4.98}$$

The transition matrices of the two systems are thus also dual to each other, according to the previous principle of duality: by transposition and time reversal.

Furthermore, with (4.97) it is possible to write:

$$\mathbf{x}^{\mathrm{T}}(t)\,\boldsymbol{\xi}(t)=\mathbf{x}^{\mathrm{T}}(t_0)\,\mathbf{\Phi}^{\mathrm{T}}(t,t_0)\,\mathbf{\Psi}(t,t_0)\,\boldsymbol{\xi}(t_0)=\mathbf{x}^{\mathrm{T}}(t_0)\,\boldsymbol{\xi}(t_0)\,.$$

In other words, the dot product of the state vectors of the two dual systems is constant.

4.7 Correspondence of Notations in this Book with those of MATLAB®

Equations (4.88) (*a*) and (*b*), associated with (4.89), once the time dependence indices of the parameter matrices have been suppressed, are those of the state representation used by MATLAB® in its script *kalman.m* to model a Kalman filter in terms of an LTI model[1]. The inputs to this linear system are respectively the input $\mathbf{u}_k$ and the measured output $\mathbf{y}_k$ of the plant whose state is to be estimated, while the filter output is constituted by the output $\widehat{\mathbf{y}}_{k|k}$ and the state $\widehat{\mathbf{x}}_{k|k}$, both estimated after the measurement.

This script solves the *stationary* Kalman filter equations. It calculates also the steady-state estimation error variance before and after measurement $\mathbf{\Sigma}^-$ and $\mathbf{\Sigma}^+$, according to (4.75) and (4.77). The correspondence between the MATLAB® notations and those of this book is summarized in Table 4.3, for the discrete systems.

Table 4.3 Correspondence of notations with MATLAB®.

MATLAB®	This book
$\mathbf{A}$	$\mathbf{\Phi}$
$\mathbf{B}$	$\mathbf{\Gamma}$
$\mathbf{w}$	$\mathbf{v}$
$\mathbf{v}$	$\mathbf{w}$
$\mathbf{G}$	$\mathbf{E}$
$\mathbf{QN}$	$\bar{\mathbf{Q}}$
$\mathbf{RN}$	$\bar{\mathbf{R}}$
$\widehat{\mathbf{x}}[n\,|\,n]$	$\widehat{\mathbf{x}}_{k|k}$
$\widehat{\mathbf{x}}[n+1\,|\,n]$	$\widehat{\mathbf{x}}_{k+1|k}$
$\mathbf{M}$	$\mathbf{K}$
$\mathbf{L}$	$\mathbf{\Phi K}$
$\mathbf{P}$	$\mathbf{\Sigma}_{k|k-1}$ or $\mathbf{\Sigma}^-$
$\mathbf{Z}$	$\mathbf{\Sigma}_{k|k}$ or $\mathbf{\Sigma}^+$

[1] The script file *kalman.m* is in the *Control Systems Toolbox*. On-line help is available by help kalman and help dkalman for the continuous and discrete Kalman filters.

4.8 Extended Kalman Filter

Though this book deals almost exclusively with linear systems, a brief exception will be made in this section. There is indeed an interesting extension of the Kalman filter to nonlinear systems, named Extended Kalman Filter (EKF). It will be presented briefly here, only in the discrete case which represents the majority of applications. Some details about this subject can be found in [Sim06], [Rib04].

4.8.1 Basic Idea

The basic idea, due to Stanley Schmidt, is the following. To reuse as much as possible the previous linear formulations, the approach will necessarily consist in linearizing the nonlinear system to be estimated, around a nominal state trajectory. But how to determine the nominal trajectory? The system being nonlinear, this might be a non trivial issue. The idea was thus to use the Kalman filter estimate of the system state as the nominal trajectory. The extended Kalman filter operates thus in sort of a bootstrap way: at each step, the nonlinear system is linearized around the Kalman filter estimate, which becomes the new nominal state, and the Kalman filter estimate of the next step is based on this linearized system.

4.8.2 Problem Presentation

Assume a nonlinear system is described by the following state equations:

$$\begin{cases} \mathbf{X}_{k+1} = \mathbf{f}_k(\mathbf{X}_k, \mathbf{u}_k, \mathbf{V}_k) \\ \mathbf{Y}_k = \mathbf{g}_k(\mathbf{X}_k, \mathbf{u}_k, \mathbf{W}_k) \end{cases}$$

where the process noise, $\mathbf{V}_k$, and measurement noise, $\mathbf{W}_k$, are supposed to be distinct white noises, thus not correlated with each other. The hypotheses are thus the same as in the linear situation (see (4.44)), and are summarized here again:

$$\mathrm{E}\{\mathbf{V}_k\} = 0, \quad \mathrm{Cov}\{\mathbf{V}_k, \mathbf{V}_j\} = \bar{\mathbf{Q}}_k\, \delta_{kj},$$

$$\mathrm{E}\{\mathbf{W}_k\} = 0, \quad \mathrm{Cov}\{\mathbf{W}_k, \mathbf{W}_j\} = \bar{\mathbf{R}}_k\, \delta_{kj},$$

$$\mathrm{Cov}\{\mathbf{V}_k, \mathbf{X}_j\} = 0, \quad \forall\, j \le k,$$

$$\mathrm{Cov}\{\mathbf{W}_k, \mathbf{X}_j\} = 0, \quad \forall\, j, k.$$

We will denote, as for the linear Kalman filter, by

- $\widehat{\mathbf{x}}_{k|k-1}$, the optimal estimate of the state $\mathbf{X}_k$ elaborated *before* the measurement $\mathbf{y}_k$, or *prediction* of the state $\mathbf{X}_k$;
- $\widehat{\mathbf{x}}_{k|k}$, the optimal estimate of $\mathbf{X}_k$ elaborated *after* the measurement $\mathbf{y}_k$, or *estimation* of the state $\mathbf{X}_k$.

To derive the equations of the discrete extended Kalman filter, we will proceed in two phases: linearization of the nonlinear equations, and then application of a linear Kalman filter.

4.8.3 Linearization of the Nonlinear Dynamic Equations

The evolution equation is linearized first, by Taylor series expansion in the vicinity of $\mathbf{X}_k = \widehat{\mathbf{x}}_{k|k}$ and of $\mathbf{V}_k = 0$, with our notations of Sect. B.3.2 (Appendix B):

$$\mathbf{X}_{k+1} = \mathbf{f}_k(\widehat{\mathbf{x}}_{k|k}, \mathbf{u}_k, 0) + \left.\frac{\partial \mathbf{f}_k(\mathbf{X}_k, \mathbf{u}_k, \mathbf{V}_k)}{\partial \mathbf{X}_k^{\mathrm{T}}}\right|_{\mathbf{X}_k = \hat{\mathbf{x}}_{k|k}} (\mathbf{X}_k - \widehat{\mathbf{x}}_{k|k}) + \left.\frac{\partial \mathbf{f}_k(\mathbf{X}_k, \mathbf{u}_k, \mathbf{V}_k)}{\partial \mathbf{V}_k^{\mathrm{T}}}\right|_{\mathbf{X}_k = \hat{\mathbf{x}}_{k|k}} \mathbf{V}_k \,. \tag{4.99}$$

Let us introduce the following simplifying notations, the derivative of a vector with respect to a vector being a matrix (see (B.12), Appendix B):

$$\mathbf{F}_k = \left.\frac{\partial \mathbf{f}_k(\mathbf{X}_k, \mathbf{u}_k, \mathbf{V}_k)}{\partial \mathbf{X}_k^{\mathrm{T}}}\right|_{\mathbf{X}_k = \hat{\mathbf{x}}_{k|k}} \quad \text{and} \quad \mathbf{L}_k = \left.\frac{\partial \mathbf{f}_k(\mathbf{X}_k, \mathbf{u}_k, \mathbf{V}_k)}{\partial \mathbf{V}_k^{\mathrm{T}}}\right|_{\mathbf{X}_k = \hat{\mathbf{x}}_{k|k}} .$$

Note that these two matrices are deterministic. (4.99) is then rewritten as follows:

$$\begin{aligned} \mathbf{X}_{k+1} &= \mathbf{f}_k(\widehat{\mathbf{x}}_{k|k}, \mathbf{u}_k, 0) + \mathbf{F}_k(\mathbf{X}_k - \widehat{\mathbf{x}}_{k|k}) + \mathbf{L}_k \mathbf{V}_k \\ &= \mathbf{F}_k \mathbf{X}_k + \left[\mathbf{f}_k(\widehat{\mathbf{x}}_{k|k}, \mathbf{u}_k, 0) - \mathbf{F}_k \widehat{\mathbf{x}}_{k|k}\right] + \mathbf{L}_k \mathbf{V}_k \,. \end{aligned}$$

By defining the following quantity, also deterministic,

$$\mathbf{s}_k = \mathbf{f}_k(\widehat{\mathbf{x}}_{k|k}, \mathbf{u}_k, 0) - \mathbf{F}_k \widehat{\mathbf{x}}_{k|k} \,,$$

we obtain finally:

$$\mathbf{X}_{k+1} = \mathbf{F}_k \mathbf{X}_k + \mathbf{s}_k + \mathbf{L}_k \mathbf{V}_k \ . \tag{4.100}$$

The measurement equation is linearized next by Taylor series expansion in the vicinity of $\mathbf{X}_k = \widehat{\mathbf{x}}_{k|k-1}$ and of $\mathbf{W}_k = 0$:

$$\mathbf{Y}_k = \mathbf{g}_k(\widehat{\mathbf{x}}_{k|k-1}, \mathbf{u}_k, 0) + \left.\frac{\partial \mathbf{g}_k(\mathbf{X}_k, \mathbf{u}_k, \mathbf{W}_k)}{\partial \mathbf{X}_k^{\mathrm{T}}}\right|_{\mathbf{X}_k = \widehat{\mathbf{x}}_{k|k-1}} (\mathbf{X}_k - \widehat{\mathbf{x}}_{k|k-1})$$
$$+ \left.\frac{\partial \mathbf{g}_k(\mathbf{X}_k, \mathbf{u}_k, \mathbf{W}_k)}{\partial \mathbf{W}_k^{\mathrm{T}}}\right|_{\mathbf{X}_k = \widehat{\mathbf{x}}_{k|k-1}} \mathbf{W}_k \ . \tag{4.101}$$

The following deterministic matrices are defined here, for simplification:

$$\mathbf{G}_k = \left.\frac{\partial \mathbf{g}_k(\mathbf{X}_k, \mathbf{u}_k, \mathbf{W}_k)}{\partial \mathbf{X}_k^{\mathrm{T}}}\right|_{\mathbf{X}_k = \widehat{\mathbf{x}}_{k|k-1}} \quad \text{and} \quad \mathbf{M}_k = \left.\frac{\partial \mathbf{g}_k(\mathbf{X}_k, \mathbf{u}_k, \mathbf{W}_k)}{\partial \mathbf{W}_k^{\mathrm{T}}}\right|_{\mathbf{X}_k = \widehat{\mathbf{x}}_{k|k-1}} .$$

(4.101) rewrites then

$$\begin{aligned}\mathbf{Y}_k &= \mathbf{g}_k(\widehat{\mathbf{x}}_{k|k-1}, \mathbf{u}_k, 0) + \mathbf{G}_k(\mathbf{X}_k - \widehat{\mathbf{x}}_{k|k-1}) + \mathbf{M}_k \mathbf{W}_k \\ &= \mathbf{G}_k \mathbf{X}_k + \left[\mathbf{g}_k(\widehat{\mathbf{x}}_{k|k-1}, \mathbf{u}_k, 0) - \mathbf{G}_k \widehat{\mathbf{x}}_{k|k-1}\right] + \mathbf{M}_k \mathbf{W}_k .\end{aligned}$$

By defining the deterministic quantity

$$\mathbf{r}_k = \mathbf{g}_k(\widehat{\mathbf{x}}_{k|k-1}, \mathbf{u}_k, 0) - \mathbf{G}_k \widehat{\mathbf{x}}_{k|k-1} ,$$

we obtain finally:

$$\mathbf{Y}_k = \mathbf{G}_k \mathbf{X}_k + \mathbf{r}_k + \mathbf{M}_k \mathbf{W}_k \ . \tag{4.102}$$

4.8.4 Derivation of the Discrete EKF Equations

We have now a plant described by a linear evolution equation and by a measurement equation, which is also linear. It is thus possible to apply a *linear* Kalman filter to this plant to estimate its state.

We will use the equations of the linear Kalman filter, with a *deterministic input* to the plant, seen in Sect. 4.3.4. The discrete linear plant being described by equations (4.79), recalled here for commodity,

$$\begin{cases} \mathbf{X}_{k+1} = \mathbf{\Phi}_k \, \mathbf{X}_k + \mathbf{\Gamma}_k \mathbf{u}_k + \mathbf{E}_k \mathbf{V}_k \\ \mathbf{Y}_k = \mathbf{C}_k \, \mathbf{X}_k + \mathbf{D}_k \mathbf{u}_k + \mathbf{W}_k \end{cases}$$

the discrete linear Kalman filter obeys the following equations, recalled also from Sect. 4.3.4:

- Estimation (or update) phase (equations (4.81), (4.82), (4.55) and (4.56)):

$$\widehat{\mathbf{x}}_{k|k} = \widehat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k (\mathbf{y}_k - \mathbf{C}_k \widehat{\mathbf{x}}_{k|k-1} - \mathbf{D}_k \mathbf{u}_k) ,$$

$$\widehat{\mathbf{y}}_{k|k} = \mathbf{C}_k \widehat{\mathbf{x}}_{k|k} + \mathbf{D}_k \mathbf{u}_k ,$$

$$\mathbf{K}_k = \mathbf{\Sigma}_{k|k-1} \mathbf{C}_k^{\mathrm{T}} (\mathbf{C}_k \mathbf{\Sigma}_{k|k-1} \mathbf{C}_k^{\mathrm{T}} + \overline{\mathbf{R}}_k)^{-1} = \mathbf{\Sigma}_{k|k} \mathbf{C}_k^{\mathrm{T}} \overline{\mathbf{R}}_k^{-1} ,$$

$$\begin{aligned} \mathrm{Var}\left\{\tilde{\mathbf{X}}_{k|k}\right\} = \mathbf{\Sigma}_{k|k} &= \mathbf{\Sigma}_{k|k-1} - \mathbf{\Sigma}_{k|k-1} \mathbf{C}_k^{\mathrm{T}} (\mathbf{C}_k \mathbf{\Sigma}_{k|k-1} \mathbf{C}_k^{\mathrm{T}} + \overline{\mathbf{R}}_k)^{-1} \mathbf{C}_k \mathbf{\Sigma}_{k|k-1} \\ &= (\mathbf{I} - \mathbf{K}_k \mathbf{C}_k) \mathbf{\Sigma}_{k|k-1} \end{aligned}$$

- Prediction (or propagation) phase (equations (4.83) and (4.59)):

$$\widehat{\mathbf{x}}_{k+1|k} = \mathbf{\Phi}_k \widehat{\mathbf{x}}_{k|k} + \mathbf{\Gamma}_k \mathbf{u}_k ,$$

$$\mathbf{\Sigma}_{k+1|k} = \mathbf{\Phi}_k \mathbf{\Sigma}_{k|k} \mathbf{\Phi}_k^{\mathrm{T}} + \mathbf{E}_k \overline{\mathbf{Q}}_k \mathbf{E}_k^{\mathrm{T}} .$$

By making in these equations the following substitutions:

$$\mathbf{\Phi}_k = \mathbf{F}_k ,$$

$$\mathbf{\Gamma}_k \mathbf{u}_k = \mathbf{s}_k ,$$

$$\mathbf{E}_k = \mathbf{L}_k ,$$

$$\mathbf{C}_k = \mathbf{G}_k ,$$

and

$$\mathbf{D}_k \mathbf{u}_k = \mathbf{r}_k ,$$

and by noting that, since the white noise $\mathbf{W}_k$ is centered, the noise $\mathbf{Z}_k = \mathbf{M}_k \mathbf{W}_k$ involved in (4.102) is also centered, and that its variance amounts to

$$\mathrm{Var}\{\mathbf{Z}_k\} = \mathrm{E}\left\{\mathbf{M}_k \mathbf{W}_k (\mathbf{M}_k \mathbf{W}_k)^{\mathrm{T}}\right\} = \mathbf{M}_k \, \mathrm{E}\left\{\mathbf{W}_k \mathbf{W}_k^{\mathrm{T}}\right\} \mathbf{M}_k^{\mathrm{T}} = \mathbf{M}_k \overline{\mathbf{R}}_k \mathbf{M}_k^{\mathrm{T}} ,$$

the following equations result for the discrete extended Kalman filter:

Estimation (or update) phase:

$$\begin{aligned}
\mathbf{r}_k &= \mathbf{g}_k(\widehat{\mathbf{x}}_{k|k-1},\mathbf{u}_k,0)-\mathbf{G}_k\widehat{\mathbf{x}}_{k|k-1} \\
\widehat{\mathbf{x}}_{k|k} &= \widehat{\mathbf{x}}_{k|k-1}+\mathbf{K}_k(\mathbf{y}_k-\mathbf{G}_k\widehat{\mathbf{x}}_{k|k-1}-\mathbf{r}_k) \\
&= \widehat{\mathbf{x}}_{k|k-1}+\mathbf{K}_k\left[\mathbf{y}_k-\mathbf{g}_k(\widehat{\mathbf{x}}_{k|k-1},\mathbf{u}_k,0)\right]=\widehat{\mathbf{x}}_{k|k-1}+\mathbf{K}_k(\mathbf{y}_k-\widehat{\mathbf{y}}_{k|k-1})
\end{aligned} \tag{4.103}$$

$$\mathbf{K}_k=\mathbf{\Sigma}_{k|k-1}\mathbf{G}_k^{\mathrm{T}}(\mathbf{G}_k\,\mathbf{\Sigma}_{k|k-1}\mathbf{G}_k^{\mathrm{T}}+\mathbf{M}_k\bar{\mathbf{R}}_k\mathbf{M}_k^{\mathrm{T}})^{-1}=\mathbf{\Sigma}_{k|k}\,\mathbf{G}_k^{\mathrm{T}}(\mathbf{M}_k\bar{\mathbf{R}}_k\mathbf{M}_k^{\mathrm{T}})^{-1} \tag{4.104}$$

$$\mathrm{Var}\left\{\tilde{\mathbf{X}}_{k|k}\right\}=\mathbf{\Sigma}_{k|k}$$

$$=\mathbf{\Sigma}_{k|k-1}-\mathbf{\Sigma}_{k|k-1}\mathbf{G}_k^{\mathrm{T}}(\mathbf{G}_k\,\mathbf{\Sigma}_{k|k-1}\mathbf{G}_k^{\mathrm{T}}+\mathbf{M}_k\bar{\mathbf{R}}_k\mathbf{M}_k^{\mathrm{T}})^{-1}\mathbf{G}_k\,\mathbf{\Sigma}_{k|k-1} \tag{4.105}$$

$$=(\mathbf{I}-\mathbf{K}_k\mathbf{G}_k)\mathbf{\Sigma}_{k|k-1} \tag{4.106}$$

Prediction (or propagation) phase:

$$\widehat{\mathbf{x}}_{k+1|k}=\mathbf{f}_k(\widehat{\mathbf{x}}_{k|k},\mathbf{u}_k,0) \tag{4.107}$$

$$\mathbf{\Sigma}_{k+1|k}=\mathbf{F}_k\,\mathbf{\Sigma}_{k|k}\,\mathbf{F}_k^{\mathrm{T}}+\mathbf{L}_k\,\bar{\mathbf{Q}}_k\,\mathbf{L}_k^{\mathrm{T}},\quad \mathbf{\Sigma}_{0|-1}\ \text{given} \tag{4.108}$$

Remark 4.14. Equation (4.103) lets us understand why the measurement equation (4.101) has been linearized around $\mathbf{X}_k=\widehat{\mathbf{x}}_{k|k-1}$ and not $\mathbf{X}_k=\widehat{\mathbf{x}}_{k|k}$. The reason is that the correction term in the update phase involves the difference between the *new measurement*, $\mathbf{y}_k$, and its prediction, $\widehat{\mathbf{y}}_{k|k-1}=\mathbf{g}_k(\widehat{\mathbf{x}}_{k|k-1},\mathbf{u}_k,0)$. If (4.101) had been linearized around $\mathbf{X}_k=\widehat{\mathbf{x}}_{k|k}$, leading to $\mathbf{r}_k=\mathbf{g}_k(\widehat{\mathbf{x}}_{k|k},\mathbf{u}_k,0)-\mathbf{G}_k\widehat{\mathbf{x}}_{k|k}$, then in the update phase $\widehat{\mathbf{x}}_{k|k}$ would be used to calculate the value of … $\widehat{\mathbf{x}}_{k|k}$ itself!

4.8.5 Condensed Form of the Estimation and Prediction Equations

- Estimator filter, obtained by eliminating $\widehat{\mathbf{x}}_{k|k-1}$ between (4.103) and (4.107):

$$\widehat{\mathbf{x}}_{k|k}=\mathbf{f}_{k-1}(\widehat{\mathbf{x}}_{k-1|k-1},\mathbf{u}_{k-1},0)+\mathbf{K}_k\left[\mathbf{y}_k-\mathbf{G}_k\mathbf{f}_{k-1}(\widehat{\mathbf{x}}_{k-1|k-1},\mathbf{u}_{k-1},0)-\mathbf{r}_k\right]. \tag{4.109}$$

- Predictor filter, obtained by eliminating $\widehat{\mathbf{x}}_{k|k}$ between the same equations:

$$\widehat{\mathbf{x}}_{k+1|k}=\mathbf{f}_k\left[\widehat{\mathbf{x}}_{k|k-1}+\mathbf{K}_k(\mathbf{y}_k-\mathbf{G}_k\widehat{\mathbf{x}}_{k|k-1}-\mathbf{r}_k),\mathbf{u}_k,0\right]. \tag{4.110}$$

4.9 Solved Exercises

Exercise 4.1 Discrete-time Kalman Filtering of an Air-flow Heater

Let us consider again the air-flow heater described in Exercise 3.6. As in Chap. 3, the present exercise will be solved entirely on a simulation model of this device. The plant being intrinsically noisy, it will be simulated by incorporating two pseudo-white noise generators in the Simulink® diagram, one for the measurement noise, the other one for the process noise.

a) Design by means of the program *MMCE.m*, for the discrete model of the process trainer PT326, without integral action, a state feedback by pole placement, by imposing to the closed loop the continuous values $s_i = -10,\ i = 1,2$. Design then a Kalman filter, by choosing for the covariance matrices of the process noise and of the measurement noise the following values: $\bar{Q} = 100$; $\bar{R} = 1$.

Simulate the responses by connecting alternatively the two noise generators to the diagram by means of the available switches, and by letting the Kalman filter operate in *free wheel* in order to see better its effect, the plant remaining fed back from its real (simulated) state vector.

b) Repeat the Kalman filter design with: $\bar{Q} = 1$; $\bar{R} = 100$.

Repeat the simulations with this time the filter included in the loop. Compare the results.

c) In order to illustrate the control-estimation duality, calculate instead of the Kalman filter an identity observer, and design it by quadratic criterion, by choosing the following weighting matrices, and then compare the results with those of the filter of question (a):

$$\mathbf{Q} = \mathbf{E} \cdot 100 \cdot \mathbf{E}^{\mathrm{T}}\text{, where } \mathbf{E} = \mathbf{\Gamma}\ ;\ \ R = 1$$

Solution:

(a) The program *MMCE.m* yields $\boldsymbol{\ell}^{\mathrm{T}} = \begin{pmatrix} -0.469 & 0.242 \end{pmatrix}$, $M = 2.465$, and, for the Kalman filter:

- steady state Kalman gain: $\mathbf{K} = \begin{pmatrix} 0.848 \\ 10.891 \end{pmatrix}$;

- matrices of the filter state equation (4.88)(a), with $\mathbf{D}_k = 0$, in the form of an identity observer, according to the equations (2.9) and (2.10):

$$\widehat{\mathbf{x}}_{k+1} = \mathbf{F}\,\widehat{\mathbf{x}}_k + \mathbf{\Gamma}\mathbf{u}_k + \mathbf{G}\mathbf{y}_k ,$$

$$\mathbf{G} = \mathbf{\Phi}\mathbf{K} = \begin{pmatrix} 1.055 \\ 7.589 \end{pmatrix}, \;\; \mathbf{F} = \mathbf{\Phi} - \mathbf{G}\mathbf{C} = \begin{pmatrix} -0.965 & 0.090 \\ -9.027 & 0.809 \end{pmatrix}.$$

Perform then a simulation, where only the process noise generator is connected to the plant. The covariance matrices chosen in this question indicate indeed that the measurements are only slightly disturbed, by comparison with the state which is it on the contrary strongly.

The left side of Fig. 4.9 shows the response of the real states to a square wave reference, whereas its right side shows the response of the filtered states, the Kalman filter remaining out of the control loop. The comparison of these plots illustrates clearly the filtering effect.

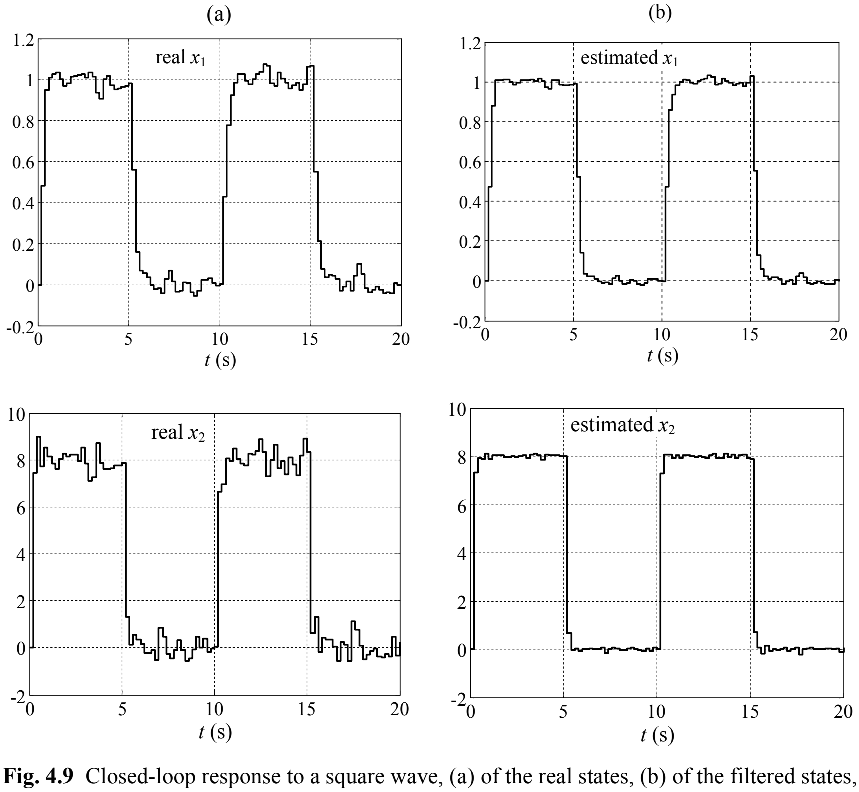


Fig. 4.9 Closed-loop response to a square wave, (a) of the real states, (b) of the filtered states, the Kalman filter being out of the loop.

If now the measurement noise generator is connected and the other one disconnected, there is no longer any filtering effect. Even worse: since the state is not disturbed at all in this case, but the measurement is, and since we “made the filter believe” that the measurement is very reliable, it puts great confidence on it to reconstruct the state and reproduces on the latter the full measurement noise!

(b) The following matrices are obtained here:

$$\mathbf{K} = \begin{pmatrix} 0.0019 \\ 0.0160 \end{pmatrix}; \ \mathbf{G} = \mathbf{\Phi K} = \begin{pmatrix} 0.0016 \\ 0.0102 \end{pmatrix}; \ \mathbf{F} = \mathbf{\Phi} - \mathbf{GC} = \begin{pmatrix} 0.088 & 0.090 \\ -1.448 & 0.809 \end{pmatrix}.$$

A big difference with the previous case is noticed here, as to the value of the Kalman gain **K**: since the filter has been informed this time that the measurement noise is significantly stronger than the process noise (or model uncertainty), it will give only very little weight to the innovation furnished at each step by the measurement and will rely primarily on the plant model to reconstruct the state. This situation is illustrated strikingly in Fig. 4.10.

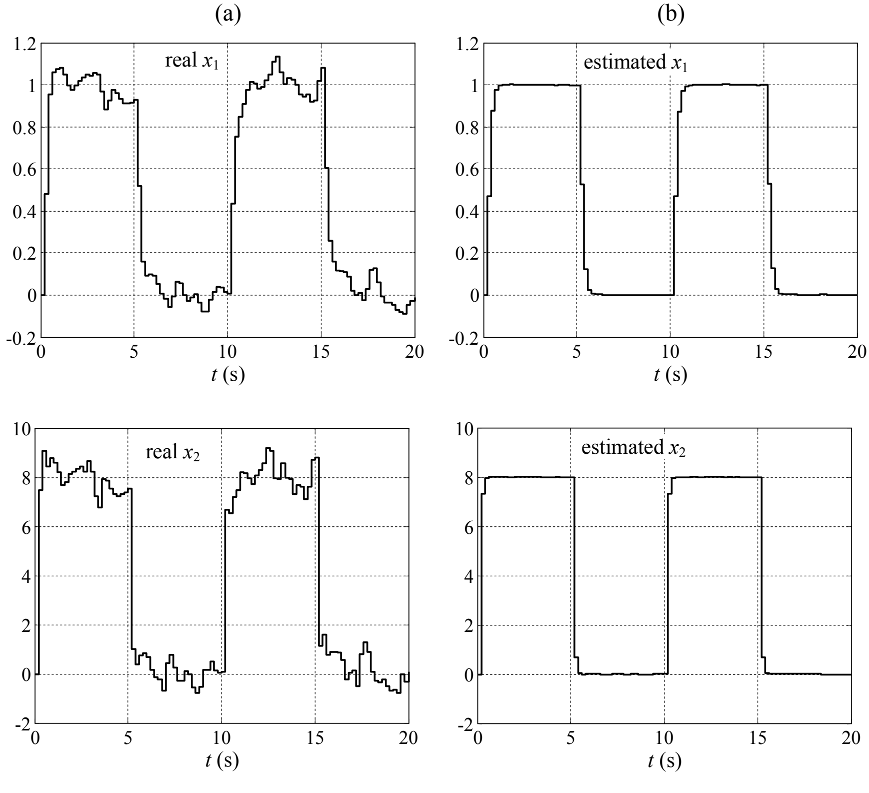


Fig. 4.10 Same as Fig. 4.9, but with filter weightings of question (b): (a) (left): real states; (b) (right): filtered states.

Here an almost perfect state variable reconstruction takes place, be it in the presence of measurement noise or of process noise. The filter operates here even better, *included in the loop*, since it receives at its input the control signal which has *already been filtered* at the previous step, which is not the case when it operates in free wheel.

(c) The matrices found for the observer are:

$$\mathbf{F} = \begin{pmatrix} -0.965 & 0.090 \\ -9.027 & 0.809 \end{pmatrix}, \quad \mathbf{G} = \begin{pmatrix} 1.055 \\ 7.589 \end{pmatrix},$$

which are exactly the values found in question (a).

This is a confirmation of the duality mentioned in this chapter: a Kalman filter is nothing else but the dual of an optimal controller. The *dlqr* algorithm solving a Riccati equation which is dual of that of an optimal filter should therefore yield the same values if it is given the transposed matrices, as discussed in Sect. 4.6.

The particular choice of the **Q** matrix results from the fact that the Riccati equation solved by the *kalman.m* module is of the form (4.63), where the term related to that matrix appears in the form $\mathbf{E}_k \overline{\mathbf{Q}}_k \mathbf{E}_k^{\mathrm{T}}$, whereas it is simply equal to **Q** in the Riccati equation (3.80) of the optimal control solved by *dlqr.m*.

Exercise 4.2 Control of a Magnetic Tape Drive

The plant studied in this exercise is sketched in Fig. 4.11.[2]

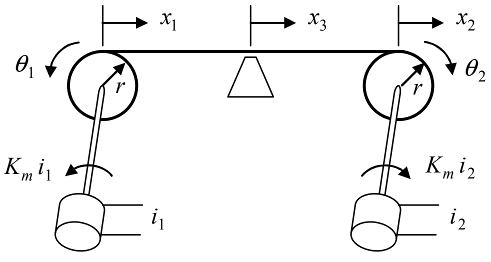


Fig. 4.11 Magnetic tape drive.

Since the magnetic tape is made of a flexible material, the objective of the control device, composed of two DC motors placed at each end of the tape and controlled separately, is to position the tape over the read head while keeping its tension *T* at some specified value. With the following parameters,

[2] Exercise inspired from [FrPW97].

$J =$ moment of inertia of a motor/capstan assembly $= 1.6 \times 10^{-4}$ kg$\cdot$m^2,

$r =$ capstan radius $= 0.1$ m ,

$K_m =$ torque constant of the motors $= 0.064$ N$\cdot$m$\cdot$A^{-1},

$k =$ coefficient of elasticity (spring) of the tape $= 2000$ N$\cdot$m^{-1},

$B =$ damping coefficient of the tape $= 3.2$ N$\cdot$m$^{-1}\cdot$s ,

the equations of motion of this system write:

$$\begin{cases} J\ddot{\theta}_1 = J\dot{\omega}_1 = -Tr + K_m i_1 \\ J\ddot{\theta}_2 = J\dot{\omega}_2 = -Tr + K_m i_2 \end{cases} \quad \text{with : } \; T = k(x_2 - x_1) + B(\dot{x}_2 - \dot{x}_1) .$$

The resulting continuous-time state model is the following:

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{\omega}_1 \\ \dot{\omega}_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 & -100 & 0 \\ 0 & 0 & 0 & 100 \\ 1.25 & -1.25 & -0.2 & -0.2 \\ 1.25 & -1.25 & -0.2 & -0.2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \omega_1 \\ \omega_2 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0.4 & 0 \\ 0 & 0.4 \end{pmatrix} \begin{pmatrix} i_1 \\ i_2 \end{pmatrix},$$

where

- $x_1, x_2 =$ tape position at the capstans (mm);
- $x_3 = (x_1 + x_2)/2 =$ tape position over read head (mm);
- $\omega_1, \omega_2 =$ angular velocities of the motor/capstan assemblies (rad/s)

The output quantities which need to be regulated, x_3 and T, are given by:

$$\begin{pmatrix} x_3 \\ T \end{pmatrix} = \begin{pmatrix} 0.5 & 0.5 & 0 & 0 \\ -2 & -2 & 0.32 & 0.32 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \omega_1 \\ \omega_2 \end{pmatrix} .$$

The specifications are the followings:

- a step change of 1 mm of the position x_3 of the tape above the read head should occur with a 1% settling time less or equal to 400 ms and an overshoot less than 10%, the tape being at rest before and after the position change;
- the tape tension T should be controlled to 2 N, while staying always between 0 and 4 N;
- the armature current of each motor must not exceed 4.9 A in steady state and 30 A during the transients.

This problem will be solved in discrete time, the plant being sampled at $T_s = 0.01$ s .

a) Suppose first the state completely accessible, and calculate an optimal state-feedback control law, with *output* weighting and the following matrices: $\mathbf{Q} = \mathbf{R} = \mathbf{I}_2$.

Determine the equivalent continuous-time poles (s_i, $i = 1,\ldots,4$) of the closed loop.

Record by simulation the step responses of x_3 and of *T*, as well as the corresponding control signals.

b) Having noticed that the specifications are not fulfilled, repeat he previous question, with the following choice: $\mathbf{Q} = 100 \times \mathbf{I}_2$ and $\mathbf{R} = \mathbf{I}_2$.

c) In order to get rid of eventual variations of plant parameters, a partial integral action is now introduced on the tension, which is the most sensitive quantity to these variations in this problem. Repeat the design of an optimal state feedback, with output weighting and $\mathbf{Q} = \begin{pmatrix} 100 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ and $\mathbf{R} = \mathbf{I}_2$, the increased size of $\mathbf{Q}$ resulting from the fact that the integrator output state variable is also taken into account in the weighting. The choice of $q_{33} = 1$ stems from the necessity not to accelerate excessively this added mode, in order to keep the control signals below the allowed limits, as will be checked in simulation.

d) Calculate an optimal estimator (Kalman filter) allowing the state reconstruction from the two only available measurements, knowing that the mean quadratic error of the position measurement amounts to an RMS value of 0.02 mm, while that of the tension measurement has an RMS value of 0.01 N. It will be assumed further that the noises and modeling errors which affect the motors are taken into account arbitrarily by the following covariance matrix of the process noise: $\bar{\mathbf{Q}} = 10^{-4} \times \mathbf{I}_2$.

Solution:

(a) Run the *MMCE.m* program by choosing the plant magnetic tape drive, discrete model, and design a state feedback, without integral action, by quadratic criterion (LQC), with output weighting, and $\mathbf{Q} = \mathbf{R} = \mathbf{I}_2$. The following feedback gain and feedforward matrices result:

$$\mathbf{L} = \begin{pmatrix} -0.534 & -0.147 & 17.32 & -1.136 \\ 0.147 & 0.534 & -1.136 & 17.32 \end{pmatrix}, \quad \mathbf{M} = \begin{pmatrix} -0.681 & 1.659 \\ 0.681 & 1.659 \end{pmatrix}.$$

The program yields the closed-loop poles in the z-plane. Their continuous-time equivalent is given by

$$s_i = (1/T_s)\ln(z_i) \quad \Rightarrow \quad s_i = -3.76 \pm j\,3.76 \text{ s}^{-1} \text{ and } -3.50 \pm j\,16.19 \text{ s}^{-1}.$$

The simulated step responses show that the specifications are not fulfilled: the 1% settling time of the position in response to a 1 mm reference step reaches 1.2 s. The control currents, on the contrary, are about 3 A in steady state, thus below the allowed nominal value.

(b) Greater weights are taken now for the trajectory, so as to accelerate the transient response by diminishing the deviation from the position step: this justifies the choice of the larger ratio between the **Q** and the **R** matrices proposed in the problem statement in this question. The control law becomes here:

$$\mathbf{L} = \begin{pmatrix} -11.4 & 5.47 & 75.40 & 19.38 \\ -5.47 & 11.74 & 19.38 & 75.40 \end{pmatrix}, \quad \mathbf{M} = \begin{pmatrix} -6.28 & 5.87 \\ 6.28 & 5.87 \end{pmatrix},$$

and the closed-loop s-poles are now:

$$s_i = -11.9 \pm j\,11.9 \text{ s}^{-1} \text{ and } -21.3 \pm j\,26.5 \text{ s}^{-1}.$$

The state-feedback gain is significantly larger than in the previous question, which provides noticeably faster closed-loop poles. This is confirmed by the step responses in Fig. 4.12.

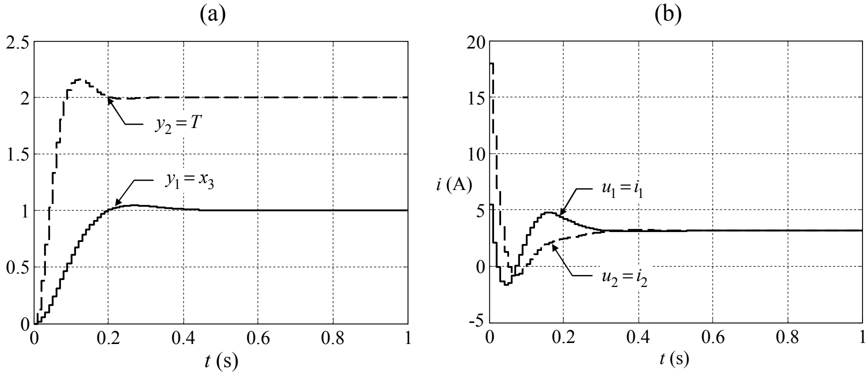


Fig. 4.12 Closed-loop step responses, without integral action: (a) outputs, (b) controls.

(c) After the model selection and the design of a state feedback, one chooses here with integral action and, for the Row-vector with indices of **y**-components to feed back through integrator(s): 2 or [2], since it is to the output $y_2 = T$ that we want to apply partial integral action. This creates the selection matrix $\mathbf{S}_i = (0 \;\; 1)$ (Sect. 1.8.3). The obtained matrices are the following:

$$\mathbf{L}_1 = \begin{pmatrix} -16.31 & 10.03 & 84.43 & 28.40 \\ -10.03 & 16.31 & 28.40 & 84.43 \end{pmatrix}; \; \mathbf{L}_2 = \begin{pmatrix} -0.547 \\ -0.547 \end{pmatrix}; \; \mathbf{M} = \begin{pmatrix} -6.28 & 8.15 \\ 6.28 & 8.15 \end{pmatrix}.$$

Note that here $\mathbf{S}_c = \mathbf{I}_2$, since $p = q$ which has allowed calculating **M** by the relation (1.16).

The step responses are given in Fig. 4.13.

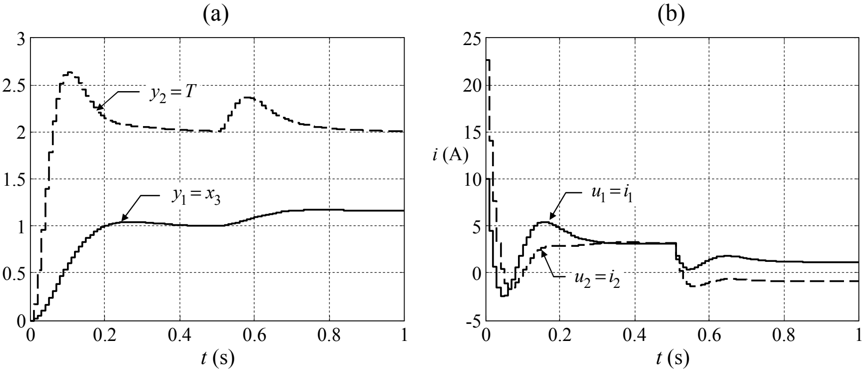


Fig. 4.13 Closed-loop responses to reference and load disturbance steps; partial integral action is applied to the tension: (a) outputs, (b) controls.

The control currents remain below 4.9 A in steady state and below 30 A during the transients. It is this last limit which would have been exceeded if, instead of $q_{33} = 1$, we would have chosen $q_{33} = 100$.

The figure shows, as in the previous case, that unit static gains are guaranteed between the two references and the corresponding outputs. Furthermore a constant load disturbance, equal to $\begin{pmatrix} 2 & 4 \end{pmatrix}^{\mathrm{T}}$ and applied through the matrix $\mathbf{E} = \mathbf{E}_m$ to the two inputs at $t = 0.5$ s, is rejected in steady state on the output $y_2 = T$, but not on $y_1 = x_3$, as expected. By repeating these simulations, the reader can verify that this is also the case for a constant measurement disturbance.

The step response overshoot of about 30% on the tension results from the zero added to the closed loop by the anticipation term **M**, namely, in the z-plane: $z_3 = 0.9328$. Since this zero is significantly "slower" than the one which the closed-loop system had without integral term and which was already an open-loop zero, $z_2 = -0.5152$, it is not surprising that it augments strongly the overshoot.

As indicated in the discussion of Sect. 1.8.3, its effect can be reduced by diminishing the gain of this term by means of a multiplicative coefficient $\rho < 1$. With $0.5 \times \mathbf{M}$, this overshoot is lowered to 15%, but at the cost of a 50% reduction of the static gain on the output y_1, which of course does not fulfill anymore the specifications.

(d) Back to the main menu choose the synthesis of a Kalman filter, with the following noise covariance matrices. For $\overline{\mathbf{R}}$, select according to the data in the problem statement and by application of (4.92):

$$\overline{\mathbf{R}} = \begin{pmatrix} 0.02^2 & 0 \\ 0 & 0.01^2 \end{pmatrix} = \begin{pmatrix} 0.0004 & 0 \\ 0 & 0.0001 \end{pmatrix}.$$

The filter is calculated obviously for the initial discrete-time plant, without integrator: it would make no sense to estimate the quantity delivered at the output of an integrator. The resulting Kalman gain is:

$$\mathbf{K} = \begin{pmatrix} 0.0518 & -0.0159 \\ 0.0518 & 0.0159 \\ -0.0014 & 0.0006 \\ 0.0014 & 0.0006 \end{pmatrix}.$$

In the simulation diagram two uncorrelated measurement noises, with respective variances of 0.02^2 mm^2 and of 0.01^2 N^2, are applied to the measured outputs of x_3 and of T made available in the diagram by the block $\mathbf{C}_m$. The obtained step responses are plotted in Fig. 4.14.

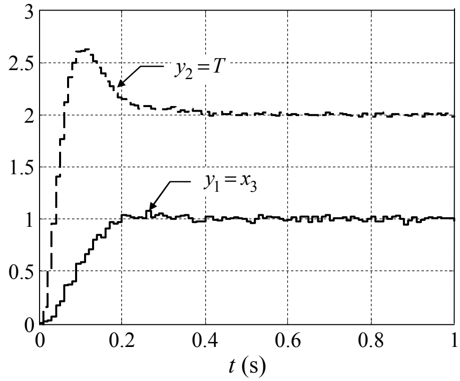


Fig. 4.14 Step responses of the two outputs in the presence of measurement noise.

The noise generator of the diagram is in fact a pseudo-white noise generator, whose power spectral density is chosen equal to $\begin{pmatrix}0.02^2 & 0.01^2\end{pmatrix} \times T_s$, in order to account for the relation between this parameter and the covariance of the signal generated by this type of block.

5 Optimal Stochastic Control

5.1 Problem Description. Discrete-time LQG Control

This chapter deals with the optimal control of a noisy linear system, the state of which is not entirely available, i.e., which requires a state reconstructor in the control loop. Since the system is submitted to random influences, a filter, e.g. an optimal filter such as the Kalman filter, will be used.

In the rather frequent case where the stochastic processes applied to the plant are Gaussian, such a control is called LQG, for *Linear Quadratic Gaussian*.

Stated like this, the problem is similar in its structure to that of a state feedback control of a deterministic system, using an observer in the loop to reconstruct part or all of its state vector.

Thus the following question arises: does the separation theorem apply here also, i.e., can we dissociate the synthesis of the LQG optimal control law from that of the Kalman filter, as we could do it in the deterministic case for the synthesis of a controller and an observer introduced in the loop?

Let us treat here the case of discrete-time systems. The plant is a noisy, time-varying, linear discrete system, described thus by equation (4.79) of Sect. 4.3.4, where however the assumption $\mathbf{D}_k = 0$ will be made to shorten the equations:

$$\begin{cases} \mathbf{X}_{k+1} = \mathbf{\Phi}_k\, \mathbf{X}_k + \mathbf{\Gamma}_k\, \mathbf{u}_k + \mathbf{E}_k \mathbf{V}_k \\ \quad \mathbf{Y}_k = \mathbf{C}_k\, \mathbf{X}_k + \mathbf{W}_k \end{cases} \tag{5.1}$$

The other hypotheses are those of Sect. 4.3.1.1: the random processes $\mathbf{V}_k$ and $\mathbf{W}_k$ are white noises, uncorrelated with each other and with the past and present states of the system, and having respective covariance matrices $\bar{\mathbf{Q}}_k$ and $\bar{\mathbf{R}}_k$. The plant initial state $\mathbf{X}_0$ is random and is described by:

$$\mathrm{E}\{\mathbf{X}_0\} = \boldsymbol{\mu}_{\mathbf{X},0}\,; \quad \mathrm{Var}\{\mathbf{X}_0\} = \mathbf{\Sigma}_0\,.$$

E. Ostertag, *Mono- and Multivariable Control and Estimation,*
DOI 10.1007/978-3-642-13734-1_5, © Springer-Verlag Berlin Heidelberg 2011

Let be given, furthermore, a quadratic cost functional of the type introduced by equation (3.72) for time-invariant discrete linear systems, but whose formulation will be extended here to the case of time-varying systems, and of weighting matrices of the two quadratic forms, $\mathbf{Q}_k$ and $\mathbf{R}_k$, eventually also time dependent:

$$\widehat{J} = \mathrm{E}\left\{\frac{1}{2}\mathbf{X}_N^{\mathrm{T}}\mathbf{S}\mathbf{X}_N + \frac{1}{2}\sum_{k=0}^{N-1}(\mathbf{X}_k^{\mathrm{T}}\mathbf{Q}_k\mathbf{X}_k + \mathbf{u}_k^{\mathrm{T}}\mathbf{R}_k\mathbf{u}_k)\right\}. \tag{5.2}$$

Note that the state vector involved in this criterion is random, which suggests to take the mathematical expectation of the expression which had been used for J in (3.72), so as to obtain a deterministic cost functional, denoted here by $\widehat{J}$.

The objective is to determine a control law $\mathbf{u}_k^*$ function of the measurements $\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_k$, which minimizes the cost functional $\widehat{J}$. This approach is called stochastic optimization of the control.

5.2 Stochastic Separation Theorem

The hypotheses are:

1. the plant is linear;
2. the cost functional is quadratic;
3. the added noises $\mathbf{V}_k$ and $\mathbf{W}_k$ are white noises, not correlated with each other, $\mathbf{V}_k$ being uncorrelated with all past and present states and $\mathbf{W}_k$ with all states.

Theorem 5.1 (separation theorem). *The optimal stochastic control of the system (5.1), with the cost functional (5.2), is obtained by taking the optimal control law* $\mathbf{u}_k = -\mathbf{L}_k\,\mathbf{x}_k$ *given by (3.78) and (3.79), calculated for the corresponding deterministic system without noises,*

$$\mathbf{x}_{k+1} = \mathbf{\Phi}_k\,\mathbf{x}_k + \mathbf{\Gamma}_k\mathbf{u}_k\,,$$

with the cost functional

$$J = \frac{1}{2}\mathbf{x}_N^{\mathrm{T}}\mathbf{S}\mathbf{x}_N + \frac{1}{2}\sum_{k=0}^{N-1}(\mathbf{x}_k^{\mathrm{T}}\mathbf{Q}_k\,\mathbf{x}_k + \mathbf{u}_k^{\mathrm{T}}\mathbf{R}_k\,\mathbf{u}_k)\,,$$

and by replacing in this control law the plant state $\mathbf{x}$ *by its optimal estimate* $\widehat{\mathbf{x}}_{k|k-1}$, *yielded by a Kalman filter:*

$$\widehat{\mathbf{x}}_{k|k-1} = \mathbf{\Phi}_{k-1}\,\widehat{\mathbf{x}}_{k-1|k-2} + \mathbf{\Phi}_{k-1}\mathbf{K}_{k-1}(\mathbf{y}_{k-1} - \mathbf{C}_{k-1}\widehat{\mathbf{x}}_{k-1|k-2})\,,$$

where

$$\mathbf{K}_k = \mathbf{\Sigma}_{k|k-1}\,\mathbf{C}_k^{\mathrm{T}}(\mathbf{C}_k\,\mathbf{\Sigma}_{k|k-1}\,\mathbf{C}_k^{\mathrm{T}} + \overline{\mathbf{R}}_k)^{-1}\,,$$

$$\mathbf{\Sigma}_{k+1|k} = \mathbf{\Phi}_k\mathbf{\Sigma}_{k|k-1}\mathbf{\Phi}_k^{\mathrm{T}} - \mathbf{\Phi}_k\mathbf{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}}(\mathbf{C}_k\,\mathbf{\Sigma}_{k|k-1}\mathbf{C}_k^{\mathrm{T}} + \overline{\mathbf{R}}_k)^{-1}\mathbf{C}_k\,\mathbf{\Sigma}_{k|k-1}\mathbf{\Phi}_k^{\mathrm{T}}$$
$$+\mathbf{E}_k\overline{\mathbf{Q}}_k\mathbf{E}_k^{\mathrm{T}}\,.$$

The sought control law is thus, according to (3.79) where eventual time dependencies of the matrices have been reintroduced:

$$\mathbf{u}_k = -\mathbf{L}_k\,\widehat{\mathbf{x}}_{k|k-1} = -\,(\mathbf{R}_k + \mathbf{\Gamma}_k^{\mathrm{T}}\mathbf{P}_{k+1}\mathbf{\Gamma}_k)^{-1}\mathbf{\Gamma}_k^{\mathrm{T}}\mathbf{P}_{k+1}\,\mathbf{\Phi}_k\,\widehat{\mathbf{x}}_{k|k-1}\ .$$

Such a synthesis is therefore decomposed in two independent problems:

- the optimization of the control of the deterministic, not noisy plant;
- the optimization of the estimation of the state **X** from the measurements **y**.

Partial proof of the theorem. By using the error estimation before measurement, $\widetilde{\mathbf{X}}_{k|k-1} = \mathbf{X}_k - \widehat{\mathbf{x}}_{k|k-1}$, introduced in Sect. 4.3.1.2, we can replace $\mathbf{X}_k$ in the quadratic form defined by $\mathbf{Q}_k$ involved in the expression of $\widehat{J}$ by $\mathbf{X}_k = \widehat{\mathbf{x}}_{k|k-1} + \widetilde{\mathbf{X}}_{k|k-1}$. This yields

$$\begin{aligned}\mathbf{X}_k^{\mathrm{T}}\mathbf{Q}_k\mathbf{X}_k &= (\widehat{\mathbf{x}}^{\mathrm{T}}_{k|k-1} + \widetilde{\mathbf{X}}^{\mathrm{T}}_{k|k-1})\,\mathbf{Q}_k\,(\widehat{\mathbf{x}}_{k|k-1} + \widetilde{\mathbf{X}}_{k|k-1})\\ &= \widehat{\mathbf{x}}^{\mathrm{T}}_{k|k-1}\,\mathbf{Q}_k\,\widehat{\mathbf{x}}_{k|k-1} + \widetilde{\mathbf{X}}^{\mathrm{T}}_{k|k-1}\,\mathbf{Q}_k\,\widetilde{\mathbf{X}}_{k|k-1} + 2\,\widehat{\mathbf{x}}^{\mathrm{T}}_{k|k-1}\,\mathbf{Q}_k\,\widetilde{\mathbf{X}}_{k|k-1}\,.\end{aligned}$$

Since $\mathbf{b}^{\mathrm{T}}\mathbf{a} = \mathrm{tr}(\mathbf{a}\mathbf{b}^{\mathrm{T}})$, according to (B.2), the last term of this sum writes also:

$$2\ \mathrm{tr}(\mathbf{Q}_k\ \widetilde{\mathbf{X}}_{k|k-1}\cdot\widehat{\mathbf{x}}^{\mathrm{T}}_{k|k-1})\,.$$

Now

$$\mathrm{E}\left\{\mathrm{tr}(\mathbf{Q}_k\ \widetilde{\mathbf{X}}_{k|k-1}\cdot\widehat{\mathbf{x}}^{\mathrm{T}}_{k|k-1})\right\} = \mathbf{Q}_k\ \mathrm{tr}\left(\mathrm{E}\left\{\widetilde{\mathbf{X}}_{k|k-1}\cdot\widehat{\mathbf{x}}^{\mathrm{T}}_{k|k-1}\right\}\right) = 0\,,$$

since the estimation error is orthogonal to the estimate. The corresponding term will thus disappear from (5.2). One would obviously prove that this is also the case of the term associated with the quadratic form defined by $\mathbf{S}$, so that (5.2) becomes:

$$\widehat{J} = \mathrm{E}\left\{\frac{1}{2}\widehat{\mathbf{x}}_{N|N-1}^{\mathrm{T}}\,\mathbf{S}\,\widehat{\mathbf{x}}_{N|N-1} + \frac{1}{2}\sum_{k=0}^{N-1}(\widehat{\mathbf{x}}_{k|k-1}^{\mathrm{T}}\,\mathbf{Q}_k\,\widehat{\mathbf{x}}_{k|k-1} + \mathbf{u}_k^{\mathrm{T}}\mathbf{R}_k\mathbf{u}_k)\right\}$$

$$+\mathrm{E}\left\{\frac{1}{2}\widetilde{\mathbf{X}}_{N|N-1}^{\mathrm{T}}\,\mathbf{S}\,\widetilde{\mathbf{X}}_{N|N-1} + \frac{1}{2}\sum_{k=0}^{N-1}\widetilde{\mathbf{X}}_{k|k-1}^{\mathrm{T}}\,\mathbf{Q}_k\,\widetilde{\mathbf{X}}_{k|k-1}\right\}.$$

The first expectation in this expression yields the part which will be minimized by a suitable choice of the sequence of control signals $\{\mathbf{u}_k\}$, since $\widehat{\mathbf{x}}_{k|k-1}$ depends on the sequence $\mathbf{u}_0,\dots,\mathbf{u}_{k-1}$. The second expectation will be minimized on the contrary when the estimate of $\mathbf{X}_k$ will be optimal.

This proves, at least heuristically, the separation of the posed problem into a deterministic optimal control problem and a stochastic optimal filtering problem. The controller and the filter can therefore be determined completely separately.

5.3 Continuous-time Systems

The situation is of course very similar to that of the discrete-time systems. Consider a time-varying linear system, with $\mathbf{D}(t)=0$, as given by (4.80):

$$\begin{cases}\dot{\mathbf{X}}(t) = \mathbf{A}(t)\mathbf{X}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{E}(t)\mathbf{V}(t)\\ \mathbf{Y}(t) = \mathbf{C}(t)\mathbf{X}(t) + \mathbf{W}(t)\end{cases} \tag{5.3}$$

The random processes $\mathbf{V}(t)$ and $\mathbf{W}(t)$ are white noises, not correlated with each other nor with $\mathbf{X}(t)$, and having respective power spectral densities $\mathbf{Q}(t)$ and $\mathbf{R}(t)$. The plant initial state $\mathbf{X}_0$ is random and described by:

$$\mathrm{E}\{\mathbf{X}_0\} = \boldsymbol{\mu}_{\mathbf{X},0}\ ;\quad \mathrm{Var}\{\mathbf{X}_0\} = \boldsymbol{\Sigma}_0\,.$$

The stochastic optimization of the control will have to minimize the following cost functional:

$$\widehat{J} = \mathrm{E}\left\{\frac{1}{2}\mathbf{X}^{\mathrm{T}}(t_f)\mathbf{S}\mathbf{X}(t_f) + \frac{1}{2}\int_0^{t_f}\left[\mathbf{X}^{\mathrm{T}}(t)\mathbf{Q}(t)\mathbf{X}(t) + \mathbf{u}^{\mathrm{T}}(t)\mathbf{R}(t)\mathbf{u}(t)\right]dt\right\}.$$

Assume that an optimal control law has been designed for the plant without noise, thus of the form

$$\mathbf{u}(t) = -\mathbf{L}(t)\mathbf{x}(t)\,,$$

with

$$\mathbf{L}(t) = \mathbf{R}^{-1}(t)\,\mathbf{B}^{\mathrm{T}}(t)\,\mathbf{P}(t)\,,$$

where $\mathbf{P}(t)$ is solution of the Riccati differential equation.

Since the state is not available, let us replace it in this control law by its optimal estimate $\widehat{\mathbf{x}}(t)$:

$$\mathbf{u}(t) = -\mathbf{L}(t)\widehat{\mathbf{x}}(t)\,,$$

and introduce the estimation error, $\widetilde{\mathbf{X}}(t) = \mathbf{X}(t) - \widehat{\mathbf{x}}(t)$:

$$\mathbf{u}(t) = -\mathbf{L}(t)\Big[\mathbf{X}(t) - \widetilde{\mathbf{X}}(t)\Big]\,.$$

The closed-loop state differential equation becomes:

$$\dot{\mathbf{X}}(t) = \mathbf{A}(t)\mathbf{X}(t) - \mathbf{B}(t)\mathbf{L}(t)\Big[\mathbf{X}(t) - \widetilde{\mathbf{X}}(t)\Big] + \mathbf{E}(t)\mathbf{V}(t)\,. \tag{5.4}$$

On the other hand, the equation giving the optimal estimate is, according to (4.86):

$$\begin{aligned}\dot{\widehat{\mathbf{x}}}(t) &= \mathbf{A}(t)\widehat{\mathbf{x}}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{K}(t)\Big[\mathbf{y}(t) - \mathbf{C}(t)\widehat{\mathbf{x}}(t)\Big] \\ &= \mathbf{A}(t)\widehat{\mathbf{x}}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{K}(t)\mathbf{C}(t)\Big[\mathbf{X}(t) - \widehat{\mathbf{x}}(t)\Big] + \mathbf{K}(t)\mathbf{W}(t)\,.\end{aligned}$$

By subtracting this equation from the first of (5.3), we get

$$\begin{aligned}\dot{\widetilde{\mathbf{X}}}(t) = \Big[\dot{\mathbf{X}}(t) - \dot{\widehat{\mathbf{x}}}(t)\Big] = \mathbf{A}(t)\Big[\mathbf{X}(t) - \widehat{\mathbf{x}}(t)\Big] + \mathbf{E}(t)\mathbf{V}(t) - \mathbf{K}(t)\mathbf{C}(t)\Big[\mathbf{X}(t) - \widehat{\mathbf{x}}(t)\Big] \\ -\mathbf{K}(t)\mathbf{W}(t)\,,\end{aligned}$$

i.e.,

$$\dot{\widetilde{\mathbf{X}}}(t) = \big[\mathbf{A}(t) - \mathbf{K}(t)\mathbf{C}(t)\big]\,\widetilde{\mathbf{X}}(t) + \mathbf{E}(t)\mathbf{V}(t) - \mathbf{K}(t)\mathbf{W}(t)\,. \tag{5.5}$$

Regroup now equations (5.4) and (5.5):

$$\begin{pmatrix}\dot{\mathbf{X}} \\ \dot{\widetilde{\mathbf{X}}}\end{pmatrix} = \begin{pmatrix}\mathbf{A} - \mathbf{BL} & \mathbf{BL} \\ \mathbf{0} & \mathbf{A} - \mathbf{KC}\end{pmatrix}\begin{pmatrix}\mathbf{X} \\ \widetilde{\mathbf{X}}\end{pmatrix} + \begin{pmatrix}\mathbf{E} & \mathbf{0} \\ \mathbf{E} & -\mathbf{K}\end{pmatrix}\begin{pmatrix}\mathbf{V} \\ \mathbf{W}\end{pmatrix}\,.$$

This writing shows clearly that the dynamics of the estimation error is completely decoupled from that of the state. The stochastic separation theorem takes thus here exactly the same form as the algebraic separation theorem, encountered for the design of a state-feedback control system containing an observer in the loop to reconstruct the state.

The functional diagram of such a control system in represented in Fig. 5.1.

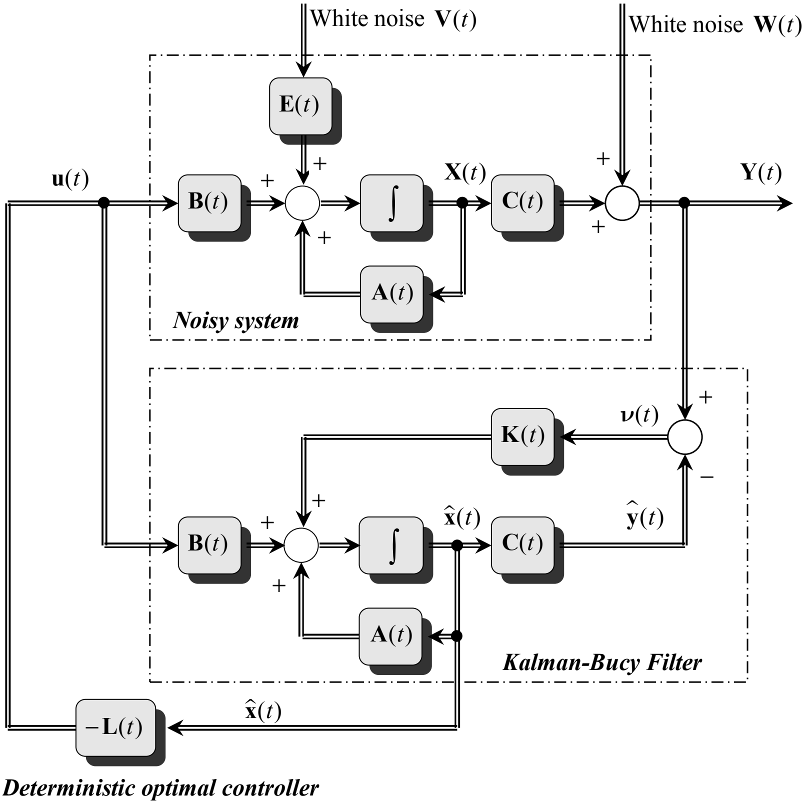


Fig. 5.1 Functional diagram of a state-feedback control system including a Kalman-Bucy filter.

5.4 Loss of Robustness in an LQG Synthesis

The inclusion of an estimator in an LQC-type control loop can lead to a loss of robustness. Since this goes beyond the scope of this book, the interested reader may consult the specialized works of this field, such as [Doy78] and [MoGB81].

An illustration of this phenomenon is proposed in an exercise at the end of this chapter.

5.5 Solved Exercise

Exercise 5.1 Discrete-time LQG Control of an Air Flow Heater

We consider again the air-flow heater used in Exercises 3.6 and 4.1. We want here to apply an LQG control to this setup.

a) Design an optimal LQC control for the discrete model of this setup, with the following weighting matrices: $\mathbf{Q} = \mathbf{I}_2$; $R = 10$.

b) Calculate a Kalman filter with the following covariance matrices of the process noise and the measurement noise: $\bar{Q} = 100$; $\bar{R} = 1$.

c) Simulate the closed loop with Simulink®, and verify its operation with the Kalman filter included in the loop.

d) Study the robustness of the obtained LQG control, by following the same approach as in question (c) of Exercise 3.6. What do you remark?

Solution:

(a) The program *MMCE.m* yields readily the following control law:

$$\boldsymbol{\ell}^{\mathrm{T}} = \begin{pmatrix} -0{,}2729 & 0{,}1427 \end{pmatrix}; \; M = 1{,}8689 \,.$$

(b) The same Kalman filter as in Exercise 4.1, question (a) is found.

(c) The responses are quite similar to those of Exercise 4.1, question (a).

(d) The script file LQ_LQG_robustness_discrete.m, which is included in the archive *mmce.zip*, plots in its second half, entitled 2) Study of the LQG control robustness (filter in the loop), the compensated open-loop Nyquist diagram and the "robustness" circle.

The statement LQGREG(KEST,L,'current') produces an LQG controller by connecting in series the Kalman filter KEST calculated by kalman.m and the state feedback matrix **L** calculated by dlqr, as sketched in Fig. 5.2.

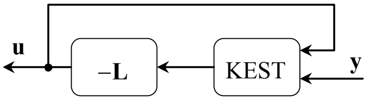


Fig. 5.2 Block diagram for the MATLAB® statements kalman.m and dlqr.

This controller is then connected in series with the plant output to build the compensated open loop, by the statement Topen_ss = series(PT326_ext, RLQG). Finally, the Nyquist diagram of the compensated open loop is calculated by the statement [Re,Im] = nyquist(Topen_ss, omega), and the *forbidden* circle is calculated according to (3.92). The obtained plot is represented in Fig. 5.3.

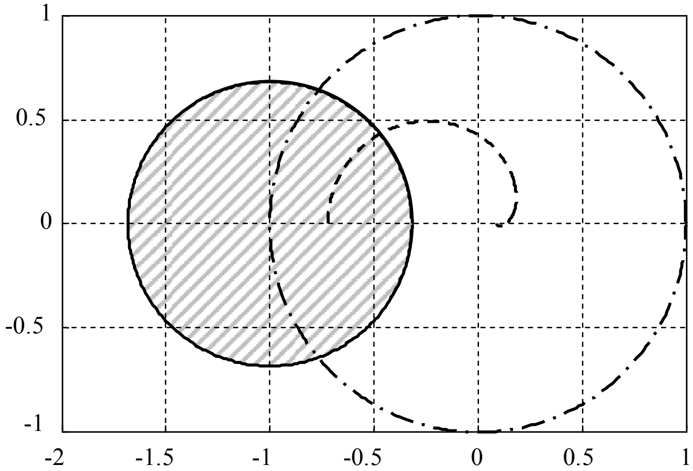


Fig. 5.3 Compensated open-loop Nyquist diagram (dashed line), forbidden disk (hatched) and unit circle (dash-dotted line).

At the difference with the plot obtained at question (c) of Exercise 3.6, it appears here that the Nyquist diagram of the compensated open loop enters the forbidden disk. The guaranteed robust stability margins are therefore lost, as a consequence of the introduction of an estimator in an LQC control loop.

6 Linear Matrix Inequalities

The origin of Linear Matrix Inequalities (LMIs) goes back as far as 1890, although they were not called this way at that time, when Lyapunov showed that the stability of a linear system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ is equivalent to the existence of a positive definite matrix **P**, which satisfies the matrix *inequality* $\mathbf{A}^{\mathrm{T}}\mathbf{P} + \mathbf{P}\mathbf{A} < 0$, expression which will be clarified below. The term “Linear Matrix Inequality” was coined by Willems in the 1970’s to refer to this specific LMI, in connection with quadratic optimal control. Due to the lack of good computers as well as of efficient algorithms to solve them, the LMIs did not receive a great deal of consideration from control and system researchers until the late 1980’s, when Nesterov and Nemirovsky developed interior-point methods that allowed solving elegantly LMI problems. New algorithms appeared then, triggering a renewed interest in this subject.

The purpose of this chapter is to give the basics of these methods and their utilization in control theory, with applications in the solved exercises at the end of the chapter. The reader can perform the solutions with the downloadable software, as is the case for all the exercises of this book.

Details about LMIs and their history can be found in books or publications, such as [BBFE93], [BEFB94], [ANAL06], and [Duc02b].

6.1 General Considerations and Main Properties

6.1.1 Definition of an LMI

A linear matrix inequality is a constraint of the form:

$$F(x) \triangleq F_0 + \sum_{i=1}^{m} x_i F_i > 0\,, \tag{6.1}$$

where

E. Ostertag, *Mono- and Multivariable Control and Estimation,*
DOI 10.1007/978-3-642-13734-1_6, © Springer-Verlag Berlin Heidelberg 2011

- $\mathbf{x} = \begin{pmatrix} x_1 & \cdots & x_m \end{pmatrix}^{\mathrm{T}} \in \mathbb{R}^m$ is the vector of the m variables,
- $\mathbf{F}_i = \mathbf{F}_i^{\mathrm{T}} \in \mathbb{R}^{n \times n}$ are given symmetric matrices,
- the inequality symbol, $>$, means that the matrix $\mathbf{F}(\mathbf{x})$ is positive definite, i.e., $\mathbf{u}^{\mathrm{T}}\mathbf{F}(x)\mathbf{u} > 0$ for all nonzero $\mathbf{u} \in \mathbb{R}^n$ (see Appendix B, Sect. B.6).

There are also nonstrict LMIs, of the form $\mathbf{F}(\mathbf{x}) \geq 0$, where $\geq$ means that the matrix $\mathbf{F}(\mathbf{x})$ is positive semidefinite, and LMIs of the form $\mathbf{F}(\mathbf{x}) < 0$ which are obviously equivalent to $-\mathbf{F}(\mathbf{x}) > 0$.

A set of constraints, or multiple LMIs, such as $\mathbf{F}^{(1)}(\mathbf{x}) > 0, \ldots, \mathbf{F}^{(p)}(\mathbf{x}) > 0$, can be grouped into a single LMI:

$$\begin{pmatrix} \mathbf{F}_1(\mathbf{x}) & 0 & \cdots & 0 \\ 0 & \mathbf{F}_2(\mathbf{x}) & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{F}_p(\mathbf{x}) \end{pmatrix} > 0 .$$

In the sequel, there will be no distinction between a set of LMIs and a single LMI.

Remark 6.1. When the matrices $\mathbf{F}_i$ are diagonal, the LMI $\mathbf{F}(\mathbf{x}) > 0$ becomes simply a set of affine inequalities.

6.1.2 Basic Properties of LMIs

Property 6.1. Given a nonsingular matrix $\mathbf{T} = \mathbf{T}^{\mathrm{T}} \in \mathbb{R}^{n \times n}$, the LMI $\mathbf{F}(\mathbf{x}) > 0$ is equivalent to $\mathbf{T}^{\mathrm{T}}\mathbf{F}(\mathbf{x})\mathbf{T} > 0$.

Proof. The proof is trivial, if one considers that the quadratic form $\mathbf{u}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\mathbf{F}(\mathbf{x})\mathbf{T}\mathbf{u}$ is the same as $\mathbf{v}^{\mathrm{T}}\mathbf{F}(\mathbf{x})\mathbf{v}$, with $\mathbf{v} = \mathbf{T}\mathbf{u}$. Therefore, if the matrix $\mathbf{F}(\mathbf{x})$ is positive definite, $\mathbf{T}^{\mathrm{T}}\mathbf{F}(\mathbf{x})\mathbf{T}$ is it also, and vice-versa.

Property 6.2: Schur Lemma

Given the following matrices: $\mathbf{Q}(\mathbf{x}) = \mathbf{Q}(\mathbf{x})^{\mathrm{T}} \in \mathbb{R}^{n \times n}$, $\mathbf{R}(\mathbf{x}) = \mathbf{R}(\mathbf{x})^{\mathrm{T}} \in \mathbb{R}^{m \times m}$, and $\mathbf{S}(\mathbf{x}) \in \mathbb{R}^{n \times m}$, which depend affinely on $\mathbf{x}$, the LMI

$$\begin{pmatrix} \mathbf{Q}(\mathbf{x}) & \mathbf{S}(\mathbf{x}) \\ \mathbf{S}(\mathbf{x})^{\mathrm{T}} & \mathbf{R}(\mathbf{x}) \end{pmatrix} > 0 \tag{6.2}$$

is equivalent to the two LMIs

$$\begin{cases} \mathbf{R}(\mathbf{x}) > 0 \\ \mathbf{Q}(\mathbf{x}) - \mathbf{S}(\mathbf{x})\mathbf{R}(\mathbf{x})^{-1}\mathbf{S}(\mathbf{x})^{\mathrm{T}} > 0 \end{cases} \tag{6.3}$$

and to the two LMIs

$$\begin{cases} \mathbf{Q}(\mathbf{x}) > 0 \\ \mathbf{R}(\mathbf{x}) - \mathbf{S}(\mathbf{x})^{\mathrm{T}}\mathbf{Q}(\mathbf{x})^{-1}\mathbf{S}(\mathbf{x}) > 0 \end{cases} \tag{6.4}$$

In other words, the set of nonlinear inequalities (6.3) or (6.4) can be represented by the LMI (6.2)

Proof. Let us prove the first equivalence, the second one being contained in the first one by simple substitution of matrices. The condition $\mathbf{R}(\mathbf{x}) > 0$ is contained implicitly in (6.2), since $\mathbf{R}(\mathbf{x})$ is one of the diagonal blocks of a positive definite matrix. With the use of Property 6.1, and by dropping temporarily the dependencies on $\mathbf{x}$ to simplify the notations, we can write:

$$\begin{pmatrix} \mathbf{I}_n & 0 \\ 0 & \mathbf{R}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{Q} & \mathbf{S} \\ \mathbf{S}^{\mathrm{T}} & \mathbf{R} \end{pmatrix} \begin{pmatrix} \mathbf{I}_n & 0 \\ 0 & \mathbf{R}^{-1} \end{pmatrix} = \begin{pmatrix} \mathbf{Q} & \mathbf{S}\mathbf{R}^{-1} \\ \mathbf{R}^{-1}\mathbf{S}^{\mathrm{T}} & \mathbf{R}^{-1} \end{pmatrix} > 0,$$

where $\mathbf{I}_n$ is the identity matrix of size $(n \times n)$. By applying again Property 6.1 to the right side of this matrix inequality, the resulting inequality

$$\begin{pmatrix} \mathbf{I}_n & -\mathbf{S} \\ 0 & \mathbf{R} \end{pmatrix} \begin{pmatrix} \mathbf{Q} & \mathbf{S}\mathbf{R}^{-1} \\ \mathbf{R}^{-1}\mathbf{S}^{\mathrm{T}} & \mathbf{R}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{I}_n & 0 \\ -\mathbf{S}^{\mathrm{T}} & \mathbf{R} \end{pmatrix} = \begin{pmatrix} \mathbf{Q} - \mathbf{S}\mathbf{R}^{-1}\mathbf{S}^{\mathrm{T}} & 0 \\ 0 & \mathbf{R} \end{pmatrix} > 0$$

is effectively the same as (6.3).

As illustrated above, this result is very useful to convert nonlinear matrix inequalities to LMI form. An interesting example is given by the following constraint:

$$\operatorname{tr}\left[\mathbf{S}(\mathbf{x})^{\mathrm{T}}\mathbf{P}(\mathbf{x})^{-1}\mathbf{S}(\mathbf{x})\right] < 1, \quad \mathbf{P}(\mathbf{x}) > 0,$$

where $\mathbf{P}(\mathbf{x}) = \mathbf{P}(\mathbf{x})^{\mathrm{T}} \in \mathbb{R}^{n \times n}$ and $\mathbf{S}(\mathbf{x}) \in \mathbb{R}^{n \times p}$ depend affinely on $\mathbf{x}$. This constraint is treated by introducing a new matrix variable $\mathbf{Z} = \mathbf{Z}^{\mathrm{T}} \in \mathbb{R}^{p \times p}$, called a *slack* variable. The previous constraint is then equivalent to the LMI in $\mathbf{x}$ and $\mathbf{Z}$:

$$\operatorname{tr}(\mathbf{Z}) < 1, \quad \begin{pmatrix} \mathbf{Z} & \mathbf{S}(\mathbf{x})^{\mathrm{T}} \\ \mathbf{S}(\mathbf{x}) & \mathbf{P}(\mathbf{x}) \end{pmatrix} > 0.$$

Matrices as variables. Many problems will be encountered in which the variables are matrices, e.g. the Lyapunov inequality already mentioned at the beginning of this chapter,

$$\mathbf{A}^{\mathrm{T}}\mathbf{P} + \mathbf{P}\mathbf{A} < 0\,, \tag{6.5}$$

associated with the stability of a linear system, having the system matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ in state representation. Here the variable, also called *decision variable* in the LMI terminology, is the symmetric matrix $\mathbf{P} = \mathbf{P}^{\mathrm{T}}$. Even though this LMI could be written explicitly in the from (6.1) by a suitable choice of $\mathbf{F}_0$ and the $\mathbf{F}_i$, it is more interesting to keep the condensed form (6.5), and to indicate which matrices are the variables in a phrase, e.g. in this case, “the LMI (6.5) in **P**”.

Another typical example is the following *quadratic* matrix inequality, or Riccati matrix inequality, which will be encountered later again:

$$\mathbf{A}^{\mathrm{T}}\mathbf{P} + \mathbf{P}\mathbf{A} + \mathbf{P}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P} + \mathbf{Q} < 0\,, \tag{6.6}$$

where **A**, **B**, $\mathbf{Q} = \mathbf{Q}^{\mathrm{T}}$, $\mathbf{R} = \mathbf{R}^{\mathrm{T}} > 0$ are given matrices of appropriate sizes, and $\mathbf{P} = \mathbf{P}^{\mathrm{T}}$ is the variable. By using the Schur Lemma, this quadratic matrix inequality in **P** can be transformed into an LMI in **P**, since it can be rewritten as the following set of two constraints:

$$\begin{cases} -\mathbf{R} < 0 \\ \mathbf{A}^{\mathrm{T}}\mathbf{P} + \mathbf{P}\mathbf{A} + \mathbf{Q} - (-\mathbf{P}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}) < 0 \end{cases}$$

i.e., with the equivalence between (6.3) and (6.2),

$$\begin{pmatrix} \mathbf{A}^{\mathrm{T}}\mathbf{P} + \mathbf{P}\mathbf{A} + \mathbf{Q} & \mathbf{P}\mathbf{B} \\ \mathbf{B}^{\mathrm{T}}\mathbf{P} & -\mathbf{R} \end{pmatrix} < 0\,.$$

Property 6.3: Elimination Lemma

Consider the following LMI in **X**, of some particular form:

$$\mathbf{Q} + \mathbf{U}\mathbf{X}\mathbf{V}^{\mathrm{T}} + \mathbf{V}\mathbf{X}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}} < 0 \tag{6.7}$$

with $\mathbf{Q} = \mathbf{Q}^{\mathrm{T}} \in \mathbb{R}^{n \times n}$, $\mathbf{U} \in \mathbb{R}^{n \times m}$, $\mathrm{rank}(\mathbf{U}) < n$, $\mathbf{V} \in \mathbb{R}^{n \times p}$, $\mathrm{rank}(\mathbf{V}) < n$, and $\mathbf{X} \in \mathbb{R}^{m \times p}$.

We intend to eliminate **X** from this LMI. Assume that $\mathbf{U}^{\perp}$ is an orthogonal complement of **U**, i.e. a matrix of size $(n-m) \times n$ such that (see Sect. B.4, in Appendix B)

$$\mathbf{U}^{\perp}\mathbf{U} = 0 \quad \text{and} \quad \operatorname{rank}\left[\mathbf{U}(\mathbf{U}^{\perp})^{\mathrm{T}}\right] = n\,, \tag{6.8}$$

and that similarly $\mathbf{V}^{\perp}$ is an orthogonal complement of **V**, i.e. a matrix of size $(n-p)\times n$ such that

$$\mathbf{V}^{\perp}\mathbf{V} = 0 \quad \text{and} \quad \operatorname{rank}\left[\mathbf{V}(\mathbf{V}^{\perp})^{\mathrm{T}}\right] = n\,. \tag{6.9}$$

The LMI (6.7) has then a solution if and only if:

$$\begin{cases} \mathbf{U}^{\perp}\mathbf{Q}(\mathbf{U}^{\perp})^{\mathrm{T}} < 0 \\ \mathbf{V}^{\perp}\mathbf{Q}(\mathbf{V}^{\perp})^{\mathrm{T}} < 0 \end{cases} \tag{6.10}$$

Proof. The proof of the *only if* part of this lemma is trivial, by left and right multiplying (6.7) successively by $\mathbf{U}^{\perp}$ and $(\mathbf{U}^{\perp})^{\mathrm{T}}$, and then by $\mathbf{V}^{\perp}$ and $(\mathbf{V}^{\perp})^{\mathrm{T}}$. The proof of the *if* part is more involved and can be found in [BEFB94].

Remark 6.2. We have limited ourselves here to the minimum properties, which will be useful for the sequel of the material presented in this chapter. There are many more properties of LMIs, which can be found e.g. in [BEFB94], [ScWe06].

6.2 Standard Problems Involving LMIs

There are two main classes of optimization problems with constraints, expressed as LMIs.

6.2.1 LMI Feasibility Problems

This problem consists in finding an $\mathbf{x} \in \mathbb{R}^m$ solution to the LMI $\mathbf{F}(\mathbf{x}) > 0$, or to determine that this LMI is infeasible, i.e. that no such **x** exists.

6.2.2 Eigenvalue Problems

The eigenvalue problem consists in minimizing the eigenvalue of a matrix, which depends affinely on some variable, subject to an LMI constraint:

$$\begin{array}{ll} \text{minimize} & \lambda \\ \text{subject to} & \lambda\mathbf{I} - \mathbf{A}(\mathbf{x}) > 0, \quad \mathbf{B}(\mathbf{x}) > 0 \end{array}$$

Such problems appear often in the equivalent form of minimizing a linear function of **x** subject to an LMI:

$$\begin{array}{ll} \text{minimize} & \mathbf{c}^{\mathrm{T}}\mathbf{x} \\ \text{subject to} & \mathbf{F}(\mathbf{x}) > 0 \end{array}$$

6.3 LMI Problems in Systems and Control

6.3.1 Stability of an Homogeneous Linear System (Lyapunov)

Given an homogeneous linear system with state vector **x**, the theory of Lyapunov claims that, if there exists a scalar function $V(\mathbf{x})$ verifying simultaneously

(i) $V(0) = 0$,
(ii) $V(\mathbf{x}) > 0, \forall \mathbf{x} \neq 0$,
(iii) $V(\mathbf{x})$ is decreasing, for any $\mathbf{x} \neq 0$,

then the equilibrium point $\mathbf{x} = 0$ is asymptotically stable, i.e. the response to an arbitrary initial condition $\mathbf{x}(0)$ tends asymptotically towards zero.

A reasonable choice for the Lyapunov function is the quadratic function $V(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{P}\mathbf{x}$, with $\mathbf{P} \in \mathbb{R}^{n \times n}, \mathbf{P} = \mathbf{P}^{\mathrm{T}} > 0$. The two first conditions are fulfilled by this function, obviously. From the positive definiteness of **P** it results that $\sqrt{\mathbf{x}^{\mathrm{T}}\mathbf{P}\mathbf{x}} = \sqrt{V(\mathbf{x})}$ is a weighted norm, which can measure the distance to the origin $\mathbf{x} = 0$ [ANAL06]. It is therefore quite natural to consider that the system is stable if the distance to the origin is always decreasing, which represents the third condition.

6.3.1.1 Continuous Case

Since in this case an homogeneous system is described by $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$, $\mathbf{x} \in \mathbb{R}^n$, the third condition expresses as

$$\begin{aligned}\dot{V}(\mathbf{x}) &= \dot{\mathbf{x}}^{\mathrm{T}}\mathbf{P}\,\mathbf{x}+\mathbf{x}^{\mathrm{T}}\mathbf{P}\dot{\mathbf{x}} = \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{P}\,\mathbf{x}+\mathbf{x}^{\mathrm{T}}\mathbf{P}\mathbf{A}\,\mathbf{x}\\ &= \mathbf{x}^{\mathrm{T}}(\mathbf{A}^{\mathrm{T}}\mathbf{P}+\mathbf{P}\mathbf{A})\mathbf{x}<0,\ \ \forall\mathbf{x}\neq\mathbf{0}\end{aligned}\quad .$$

Therefore, the equilibrium point $\mathbf{x}=0$ of this system is asymptotically stable if there exists a matrix $\mathbf{P}=\mathbf{P}^{\mathrm{T}}\in\mathbb{R}^{n\times n}$ satisfying simultaneously the two LMIs

$$\begin{cases}\mathbf{P}>0\\ \mathbf{A}^{\mathrm{T}}\mathbf{P}+\mathbf{P}\mathbf{A}<0\end{cases} \tag{6.11}$$

Remark 6.3. Though not fundamental, it is interesting to recognize that this is indeed an LMI, as defined in (6.1). By defining $\mathbf{P}_{ij}$ as the matrix whose elements are all zero except elements (i,j) and (j,i) which are equal to one, we can write $\mathbf{P}=\sum_{i=1}^{n}\sum_{j=1}^{i}\alpha_{ij}\mathbf{P}_{ij}$, and the two previous inequalities become:

$$\sum_{i=1}^{n}\sum_{j=1}^{i}\alpha_{ij}\begin{pmatrix}\mathbf{P}_{ij} & 0\\ 0 & -\mathbf{A}^{\mathrm{T}}\mathbf{P}_{ij}-\mathbf{P}_{ij}\mathbf{A}\end{pmatrix}>0\,,$$

which is indeed an LMI with a vector **x** including all the coefficients α_{ij}. The writing (6.11) in terms of an LMI in the *matrix* variable **P** is of course much more concise, and justifies the use of matrix variables as introduced in Sect. 6.1.2.

6.3.1.2 Discrete Case

An homogeneous linear system being described here by $\mathbf{x}_{k+1}=\mathbf{\Phi}\,\mathbf{x}_k$, $\mathbf{x}\in\mathbb{R}^n$, the third Lyapunov condition writes

$$\begin{aligned}V(\mathbf{x}_{k+1})-V(\mathbf{x}_k) &= \mathbf{x}_{k+1}^{\mathrm{T}}\,\mathbf{P}\,\mathbf{x}_{k+1}-\mathbf{x}_k^{\mathrm{T}}\,\mathbf{P}\,\mathbf{x}_k = \mathbf{x}_k^{\mathrm{T}}\,\mathbf{\Phi}^{\mathrm{T}}\mathbf{P}\,\mathbf{\Phi}\,\mathbf{x}_k-\mathbf{x}_k^{\mathrm{T}}\,\mathbf{P}\,\mathbf{x}_k\\ &= \mathbf{x}_k^{\mathrm{T}}(\mathbf{\Phi}^{\mathrm{T}}\mathbf{P}\,\mathbf{\Phi}-\mathbf{P})\mathbf{x}_k<0,\ \ \forall\mathbf{x}_k\neq 0\end{aligned}$$

Therefore, the equilibrium point $\mathbf{x}=0$ of this system is asymptotically stable if there exists a matrix $\mathbf{P}=\mathbf{P}^{\mathrm{T}}\in\mathbb{R}^{n\times n}$ satisfying simultaneously the two LMIs

$$\begin{cases}\mathbf{P}>0\\ \mathbf{\Phi}^{\mathrm{T}}\mathbf{P}\,\mathbf{\Phi}-\mathbf{P}<0\end{cases}. \tag{6.12}$$

6.3.2 Stabilizability and Stabilization by State Feedback

6.3.2.1 Continuous Case

Consider the system defined by $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu}$. Does there exist for this system a state feedback $\mathbf{u} = -\mathbf{Lx}$ (see (1.2)) such that the closed loop system is stable?

To answer this question, let us apply the above Lyapunov theory. Since according to (1.3) the homogeneous closed-loop system obeys the state equation $\dot{\mathbf{x}} = (\mathbf{A} - \mathbf{BL})\mathbf{x}$, this system is stable, according to (6.11), if there exists a symmetric matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$ such that:

$$\begin{cases} \mathbf{P} > 0 \\ (\mathbf{A} - \mathbf{BL})^{\mathrm{T}} \mathbf{P} + \mathbf{P}(\mathbf{A} - \mathbf{BL}) < 0 \end{cases} \tag{6.13}$$

where the decision variables are **P** and **L**. Note that the second matrix inequality is not linear, thus not an LMI, since it contains products of these two matrices.

We will now apply a procedure, which will be used many times in the sequel. First, noting that the constraint $\mathbf{P} > 0$ guarantees that $\mathbf{P}^{-1}$ exists and is nonsingular, multiply left and right the two inequalities (6.13), according to Property 6.1, by $\mathbf{P}^{-1}$ and its transpose, $(\mathbf{P}^{-1})^{\mathrm{T}} = \mathbf{P}^{-1}$:

$$\begin{cases} \mathbf{P}^{-1}\mathbf{P}\mathbf{P}^{-1} > 0, \ \textit{i.e.}\text{:}\ \mathbf{P}^{-1} > 0 \\ \mathbf{P}^{-1}(\mathbf{A} - \mathbf{BL})^{\mathrm{T}} + (\mathbf{A} - \mathbf{BL})\mathbf{P}^{-1} < 0, \\ \qquad \textit{i.e.}\text{:}\ \mathbf{P}^{-1}\mathbf{A}^{\mathrm{T}} + \mathbf{AP}^{-1} - \mathbf{P}^{-1}\mathbf{L}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}} - \mathbf{BLP}^{-1} < 0 \end{cases}$$

With the simple change of variables $\mathbf{Y} = \mathbf{P}^{-1}$ and $\mathbf{W} = \mathbf{LP}^{-1} = \mathbf{LY}$, the following LMIs in **Y** and **W** are now obtained, after some rearrangement:

$$\begin{cases} \mathbf{Y} > 0 \\ \mathbf{AY} + \mathbf{YA}^{\mathrm{T}} - \mathbf{BW} - \mathbf{W}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}} < 0 \end{cases} \tag{6.14}$$

If these LMIs have a solution (i.e. are feasible), a state feedback gain which stabilizes the system is given by $\mathbf{L} = \mathbf{W}\mathbf{Y}^{-1}$.

α-Stabilization by state feedback. If in addition to the stabilization requirement it is desired that the closed-loop eigenvalues have all a real part inferior to $-\alpha < 0$, the closed-loop transient response will decay faster than $e^{-\alpha t}$. The closed loop is then said to be α-stable, or also to have exponential stability.

It suffices to note that the α-stability of the closed-loop system $\dot{\mathbf{x}} = (\mathbf{A} - \mathbf{BL})\mathbf{x}$ is equivalent to the stability of the closed-loop system $\dot{\mathbf{x}} = (\mathbf{A}' - \mathbf{BL})\mathbf{x}$, in which the matrix $\mathbf{A}' = \mathbf{A} + \alpha\mathbf{I}_n$ is obtained from $\mathbf{A}$ by shifting the real part of its eigenvalues. Indeed, since $\det(s\mathbf{I}_n - \mathbf{A}') = \det\left[(s-\alpha)\mathbf{I}_n - \mathbf{A}\right] = \det(\sigma\mathbf{I}_n - \mathbf{A})$, where $\sigma = s - \alpha$, the eigenvalues of $\mathbf{A} - \mathbf{BL}$ will all have a real part inferior to $-\alpha$ if all those of $\mathbf{A}' - \mathbf{BL}$ have a negative real part.

From (6.14) it follows then that a matrix $\mathbf{L}$ solution of the present problem is $\mathbf{L} = \mathbf{W}\mathbf{Y}^{-1}$, with $\mathbf{Y}$ and $\mathbf{W}$ solution of the LMI

$$\begin{cases} \mathbf{Y} > 0 \\ (\mathbf{A} + \alpha\mathbf{I}_n)\mathbf{Y} + \mathbf{Y}(\mathbf{A} + \alpha\mathbf{I}_n)^{\mathrm{T}} - \mathbf{BW} - \mathbf{W}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}} < 0 \end{cases} \tag{6.15}$$

Decay rate of a continuous-time system. The decay rate of a system $\dot{\mathbf{x}} = \mathbf{A}\,\mathbf{x}$ is the largest value of $\alpha > 0$ for which this system is α-stable. In other words, $-\alpha$ is the real part of its rightmost eigenvalue. To find this value, it would be enough to solve the following optimization problem:

$$\begin{aligned} &\text{minimize } -\alpha \\ &\text{subject to} \\ &\mathbf{A}^{\mathrm{T}}\mathbf{P} + \mathbf{PA} < -2\alpha\mathbf{P}, \quad \mathbf{P} > 0, \quad \alpha > 0 \end{aligned} \tag{6.16}$$

Unfortunately, because of the product $\alpha\mathbf{P}$, this is not an LMI, and (6.16) can be solved only by dichotomy methods or with *bilinear* matrix inequality solvers, which goes beyond the scope of this chapter.

6.3.2.2 Discrete Case

Applying a state feedback control $\mathbf{u}_k = -\mathbf{L}\,\mathbf{x}_k$ to a discrete-time linear system $\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}\mathbf{u}_k$ yields the following homogeneous closed-loop system: $\mathbf{x}_{k+1} = (\mathbf{\Phi} - \mathbf{\Gamma}\mathbf{L})\mathbf{x}_k$ (see (1.13)). According to (6.12), this system is stable if there exists a symmetric matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$ such that:

$$\begin{cases} \mathbf{P} > 0 \\ (\mathbf{\Phi} - \mathbf{\Gamma}\mathbf{L})^{\mathrm{T}}\mathbf{P}(\mathbf{\Phi} - \mathbf{\Gamma}\mathbf{L}) - \mathbf{P} < 0 \end{cases}$$

Using the Schur Lemma of (6.3) and (6.2), these two matrix inequalities are equivalent to the following single matrix inequality:

$$\begin{pmatrix} -\mathbf{P} & (\mathbf{\Phi}-\mathbf{\Gamma L})^{\mathrm{T}} \\ \mathbf{\Phi}-\mathbf{\Gamma L} & -\mathbf{P}^{-1} \end{pmatrix} < 0 . \tag{6.17}$$

This inequality is not an LMI, because of the simultaneous presence of terms in $\mathbf{P}$ and $\mathbf{P}^{-1}$. By using again Property 6.1, but pre-multiplying and post-multiplying now this inequality by the symmetric, nonsingular matrix $\begin{pmatrix} \mathbf{P}^{-1} & 0 \\ 0 & \mathbf{I}_n \end{pmatrix}$, we obtain:

$$\begin{pmatrix} -\mathbf{P}^{-1} & \mathbf{P}^{-1}(\mathbf{\Phi}-\mathbf{\Gamma L})^{\mathrm{T}} \\ (\mathbf{\Phi}-\mathbf{\Gamma L})\mathbf{P}^{-1} & -\mathbf{P}^{-1} \end{pmatrix} < 0 .$$

With the same change of variables as above, namely $\mathbf{Y}=\mathbf{P}^{-1}$ and $\mathbf{W}=\mathbf{LY}$, the following LMI in $\mathbf{Y}$ and $\mathbf{W}$ results finally:

$$\begin{pmatrix} -\mathbf{Y} & \mathbf{Y\Phi}^{\mathrm{T}}-\mathbf{W}^{\mathrm{T}}\mathbf{\Gamma}^{\mathrm{T}} \\ \mathbf{\Phi Y}-\mathbf{\Gamma W} & -\mathbf{Y} \end{pmatrix} < 0 \ . \tag{6.18}$$

If this LMI has a solution, i.e. is feasible, the state-feedback gain matrix $\mathbf{L}=\mathbf{WY}^{-1}$ stabilizes the closed-loop system.

α-Stabilization by state feedback. A discrete-time system is said to be α-stable, or also to have exponential stability, if all its eigenvalues are contained within a disk of radius $1/\alpha$, with $\alpha > 1$.

We note that the α-stability of the closed-loop system $\mathbf{x}_{k+1}=(\mathbf{\Phi}-\mathbf{\Gamma L})\,\mathbf{x}_k$ is equivalent to the stability of the closed-loop system $\mathbf{x}_{k+1}=(\mathbf{\Phi}'-\mathbf{\Gamma L})\,\mathbf{x}_k$, in which the matrix $\mathbf{\Phi}'=\alpha\mathbf{\Phi}$ is obtained from $\mathbf{\Phi}$ by multiplying its eigenvalues by α. Indeed, since $\det(z\mathbf{I}_n-\mathbf{\Phi}')=\det\left[\alpha\left(\frac{z}{\alpha}\mathbf{I}_n-\mathbf{\Phi}\right)\right]=\alpha^n\det(\xi\mathbf{I}_n-\mathbf{\Phi})$, where $\xi=z/\alpha$, the eigenvalues of $\mathbf{\Phi}-\mathbf{\Gamma L}$ will all have a modulus inferior to $1/\alpha$ if all those of $\mathbf{\Phi}'-\mathbf{\Gamma L}$ are within the unit disk centered at the origin.

From (6.18) it follows then that a matrix $\mathbf{L}$ solution of the present problem is $\mathbf{L}=\mathbf{WY}^{-1}$, with $\mathbf{Y}$ and $\mathbf{W}$ solution of the LMI

$$\begin{pmatrix} -\mathbf{Y} & \alpha\mathbf{Y\Phi}^{\mathrm{T}}-\mathbf{W}^{\mathrm{T}}\mathbf{\Gamma}^{\mathrm{T}} \\ \alpha\mathbf{\Phi Y}-\mathbf{\Gamma W} & -\mathbf{Y} \end{pmatrix} < 0 . \tag{6.19}$$

Decay rate of a discrete-time system. The decay rate of a system $\mathbf{x}_{k+1} = \mathbf{\Phi x}_k$ is the largest value of $\alpha > 1$ for which this system is α-stable. In other words, $1/\alpha$ is the modulus of its eigenvalue closest to the unit circle.

6.3.3 Detectability and Detection of Linear Systems by State Observer

The notion of detectability is dual to that of stabilizability (see Appendix A, Sect. A.3.3), exactly as observability is dual to controllability, as was shown in Sect. 2.2.2 and 4.6.

We claim that the following results are true.

Continuous case. If the following LMIs in **Y** and **V**,

$$\begin{cases} \mathbf{Y} > 0 \\ \mathbf{A}^{\mathrm{T}}\mathbf{Y} + \mathbf{YA} - \mathbf{VC} - \mathbf{C}^{\mathrm{T}}\mathbf{V}^{\mathrm{T}} < 0 \end{cases} \tag{6.20}$$

are feasible, the observer with system matrix $\mathbf{A} - \mathbf{GC}$, where $\mathbf{G} = \mathbf{Y}^{-1}\mathbf{V}$ "detects" the plant state.

Discrete case. If the following LMI in **Y** and **V**,

$$\begin{pmatrix} -\mathbf{Y} & \mathbf{Y\Phi} - \mathbf{VC} \\ \mathbf{\Phi}^{\mathrm{T}}\mathbf{Y} - \mathbf{C}^{\mathrm{T}}\mathbf{V}^{\mathrm{T}} & -\mathbf{Y} \end{pmatrix} < 0 \ , \tag{6.21}$$

is feasible, the observer with system matrix $\mathbf{\Phi} - \mathbf{G\Gamma}$, where $\mathbf{G} = \mathbf{Y}^{-1}\mathbf{V}$, "detects" the plant state.

Proof. These LMIs are readily obtained by replacing $(\mathbf{A} - \mathbf{BL})$ by $(\mathbf{A}^{\mathrm{T}} - \mathbf{C}^{\mathrm{T}}\mathbf{G}^{\mathrm{T}})$ in (6.13) for the continuous case, or $(\mathbf{\Phi} - \mathbf{\Gamma L})$ by $(\mathbf{\Phi}^{\mathrm{T}} - \mathbf{C}^{\mathrm{T}}\mathbf{G}^{\mathrm{T}})$ in (6.17) for the discrete case, and following then the same procedure as in Sect. 6.3.2 with the change of variables $\mathbf{Y} = \mathbf{P}^{-1}$ and $\mathbf{V} = \mathbf{YG}$. It may be interesting to note that the duality can be extended, in saying that to **W** of the stabilizability problem corresponds $\mathbf{V}^{\mathrm{T}} = \mathbf{G}^{\mathrm{T}}\mathbf{Y}$ of the detectability problem. This would have allowed to write (6.20) and (6.21) directly by duality from (6.14) and (6.18).

6.3.4 LQC Problem Solved with LMIs

6.3.4.1 Continuous Case

The linear quadratic control problem, which has been discussed in Sect. 3.5.6, is the following. Given a linear system, defined by

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t),\ \ \mathbf{A} \in \mathbb{R}^{n\times n},\ \ \mathbf{B} \in \mathbb{R}^{n\times p},$$

with $\mathbf{x}(t_0) = \mathbf{x}_0$ specified, find a state feedback $\mathbf{u}(t) = -\mathbf{L}\mathbf{x}(t)$ which minimizes the following quadratic criterion:

$$J = \int_0^\infty \left[\mathbf{x}^{\mathrm{T}}(t)\,\mathbf{Q}\,\mathbf{x}(t) + \mathbf{u}^{\mathrm{T}}(t)\,\mathbf{R}\,\mathbf{u}(t)\right] dt\,,$$

where $\mathbf{Q} = \mathbf{H}^{\mathrm{T}}\mathbf{H} \geq 0$ and $\mathbf{R} = \mathbf{R}^{\mathrm{T}} > 0$, with $\mathbf{H} \in \mathbb{R}^{q\times n}$, $q = \mathrm{rank}(\mathbf{Q})$, and $\mathbf{R} \in \mathbb{R}^{p\times p}$, as seen in Sect. 3.5.6, apart from the factor 1/2 which is unimportant as already mentioned in that section.

We saw in that section that the solution is obtained from the unique positive semidefinite solution of the algebraic Riccati equation (3.54).

In the following, we will be looking for a state feedback $\mathbf{u}(t) = -\mathbf{L}\,\mathbf{x}(t)$ which guarantees that the criterion J is inferior to some given number γ [Duc02b].

Since $\mathbf{u}^{\mathrm{T}}\mathbf{R}\,\mathbf{u} = \mathbf{x}^{\mathrm{T}}\mathbf{L}^{\mathrm{T}}\mathbf{R}\,\mathbf{L}\,\mathbf{x}$, let us introduce the function $V(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{P}\mathbf{x}$, with $\mathbf{P} = \mathbf{P}^{\mathrm{T}} > 0$, satisfying the two following conditions:

$$\begin{cases} V(\mathbf{x}_0) < \gamma \\ \dot{V}(\mathbf{x}) + \mathbf{x}^{\mathrm{T}}\mathbf{Q}\,\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{L}^{\mathrm{T}}\mathbf{R}\,\mathbf{L}\,\mathbf{x} < 0 \end{cases} \tag{6.22}$$

Such a function $V(\mathbf{x})$ is a Lyapunov function, since it satisfies all three conditions of Sect. 6.3.1. Furthermore,

$$\int_0^\infty \dot{V}(\mathbf{x})\,dt + \int_0^\infty (\mathbf{x}^{\mathrm{T}}\mathbf{Q}\,\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{L}^{\mathrm{T}}\mathbf{R}\,\mathbf{L}\,\mathbf{x})\,dt < 0\,, \tag{6.23}$$

which we can rewrite as

$$\underbrace{\int_0^\infty (\mathbf{x}^{\mathrm{T}}\mathbf{Q}\,\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{L}^{\mathrm{T}}\mathbf{R}\,\mathbf{L}\,\mathbf{x})\,dt}_{J} < V(\mathbf{x}_0) = \mathbf{x}_0^{\mathrm{T}}\mathbf{P}\mathbf{x}_0 < \gamma\,. \tag{6.24}$$

If there exist a matrix **L** and a function $V(\mathbf{x})$ satisfying (6.22), then **L** solves this problem. By recalling that $\dot{V}(\mathbf{x}) = \dot{\mathbf{x}}^{\mathrm{T}}\mathbf{P}\,\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{P}\dot{\mathbf{x}}$ and taking into account the closed-loop state equation, these two inequalities are equivalent to

$$\begin{cases} \mathbf{x}_0^{\mathrm{T}}\mathbf{P}\,\mathbf{x}_0 < \gamma \\ (\mathbf{A}-\mathbf{B}\mathbf{L})^{\mathrm{T}}\mathbf{P} + \mathbf{P}(\mathbf{A}-\mathbf{B}\mathbf{L}) + \mathbf{Q} + \mathbf{L}^{\mathrm{T}}\mathbf{R}\,\mathbf{L} < 0 \end{cases} \tag{6.25}$$

By left and right multiplying the second inequality by $\mathbf{Y} = \mathbf{P}^{-1}$ and introducing $\mathbf{W} = \mathbf{L}\mathbf{Y}$, we obtain successively the following inequalities:

$$\mathbf{Y}(\mathbf{A}-\mathbf{B}\mathbf{L})^{\mathrm{T}} + (\mathbf{A}-\mathbf{B}\mathbf{L})\mathbf{Y} + \mathbf{Y}\mathbf{Q}\mathbf{Y} + \mathbf{Y}\mathbf{L}^{\mathrm{T}}\mathbf{R}\,\mathbf{L}\mathbf{Y} < 0\,,$$

$$\mathbf{Y}\mathbf{A}^{\mathrm{T}} + \mathbf{A}\mathbf{Y} - \mathbf{B}\mathbf{W} - \mathbf{W}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}} + \mathbf{Y}\mathbf{H}^{\mathrm{T}}\mathbf{H}\mathbf{Y} + \mathbf{W}^{\mathrm{T}}\mathbf{R}\mathbf{W} < 0\,,$$

$$\mathbf{Y}\mathbf{A}^{\mathrm{T}} + \mathbf{A}\mathbf{Y} - \mathbf{B}\mathbf{W} - \mathbf{W}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}} + \begin{pmatrix} \mathbf{Y}\mathbf{H}^{\mathrm{T}} & \mathbf{W}^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} \mathbf{I}_n & 0 \\ 0 & \mathbf{R} \end{pmatrix} \begin{pmatrix} \mathbf{H}\mathbf{Y} \\ \mathbf{W} \end{pmatrix} < 0\,.$$

By applying the Schur Lemma, this inequality becomes the following LMI:

$$\begin{pmatrix} \mathbf{Y}\mathbf{A}^{\mathrm{T}} + \mathbf{A}\mathbf{Y} - \mathbf{B}\mathbf{W} - \mathbf{W}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}} & \mathbf{Y}\mathbf{H}^{\mathrm{T}} & \mathbf{W}^{\mathrm{T}} \\ \mathbf{H}\mathbf{Y} & -\mathbf{I}_n & 0 \\ \mathbf{W} & 0 & -\mathbf{R}^{-1} \end{pmatrix} < 0\,. \tag{6.26}$$

The first of the inequalities (6.25) becomes, successively:

$$\gamma - \mathbf{x}_0^{\mathrm{T}}\mathbf{P}\mathbf{x}_0 > 0\,,$$

$$\gamma - \mathbf{x}_0^{\mathrm{T}}\mathbf{Y}^{-1}\mathbf{x}_0 > 0\,,$$

$$\begin{pmatrix} \gamma & \mathbf{x}_0^{\mathrm{T}} \\ \mathbf{x}_0 & \mathbf{Y} \end{pmatrix} > 0\,, \tag{6.27}$$

with the use of the Schur Lemma again. Note that the initial constraint $\mathbf{P} > 0$, i.e. $\mathbf{Y} > 0$, is contained in (6.27). A state feedback matrix **L**, solution of the problem, is thus obtained by solving the LMIs (6.26) and (6.27), and letting $\mathbf{L} = \mathbf{W}\,\mathbf{Y}^{-1}$.

In order to get rid of the knowledge of $\mathbf{x}_0$, the condition (6.27) can be replaced by the condition $\mathbf{P} - \gamma\mathbf{I}_n < 0$. This guarantees that, for any $\mathbf{x}_0$, $J < \mathbf{x}_0^{\mathrm{T}}\mathbf{P}\mathbf{x}_0 < \gamma\,\mathbf{x}_0^{\mathrm{T}}\mathbf{x}_0$. The previous condition becomes then $\gamma\mathbf{I}_n - \mathbf{Y}^{-1} > 0$, i.e.,

$$\begin{pmatrix} \gamma \mathbf{I}_n & \mathbf{I}_n \\ \mathbf{I}_n & \mathbf{Y} \end{pmatrix} > 0 \ . \tag{6.28}$$

If it is desired to minimize the value of γ, the following optimization problem, which is a typical eigenvalue problem as defined in Sect. 6.2.2, should be solved:

$$\begin{gathered} \min_{\gamma, \mathbf{Y}=\mathbf{Y}^{\mathrm{T}}, \mathbf{W}} \gamma \\ \text{subject to } \ (6.26), (6.28). \end{gathered} \tag{6.29}$$

6.3.4.2 Discrete Case

Given a discrete-time system described by

$$\mathbf{x}_{k+1} = \mathbf{\Phi} \mathbf{x}_k + \mathbf{\Gamma} \mathbf{u}_k, \ \ \mathbf{\Phi} \in \mathbb{R}^{n \times n}, \ \ \mathbf{\Gamma} \in \mathbb{R}^{n \times p} ,$$

and a quadratic criterion

$$J = \sum_{k=0}^{\infty} (\mathbf{x}_k^{\mathrm{T}} \mathbf{Q}\, \mathbf{x}_k + \mathbf{u}_k^{\mathrm{T}} \mathbf{R}\, \mathbf{u}_k) ,$$

the goal is to minimize J by means of a state feedback $\mathbf{u}_k = -\mathbf{L}\, \mathbf{x}_k$, all other hypotheses being the same as in the continuous case. Let us choose here a function $V(\mathbf{x}_k) = \mathbf{x}_k^{\mathrm{T}} \mathbf{P}\, \mathbf{x}_k$ satisfying the two following conditions:

$$\begin{cases} V(\mathbf{x}_0) < \gamma \\ V(\mathbf{x}_{k+1}) - V(\mathbf{x}_k) + \mathbf{x}_k^{\mathrm{T}} \mathbf{Q}\, \mathbf{x}_k + \mathbf{x}_k^{\mathrm{T}} \mathbf{L}^{\mathrm{T}} \mathbf{R}\, \mathbf{L}\, \mathbf{x}_k < 0 \end{cases}$$

The inequalities (6.23) and (6.24) are replaced by

$$\sum_{k=0}^{\infty} \left[V(\mathbf{x}_{k+1}) - V(\mathbf{x}_k) \right] + \sum_{k=0}^{\infty} (\mathbf{x}_k^{\mathrm{T}} \mathbf{Q}\, \mathbf{x}_k + \mathbf{x}_k^{\mathrm{T}} \mathbf{L}^{\mathrm{T}} \mathbf{R}\, \mathbf{L}\, \mathbf{u}_k) < 0 ,$$

$$\underbrace{\sum_{k=0}^{\infty} (\mathbf{x}_k^{\mathrm{T}} \mathbf{Q}\, \mathbf{x}_k + \mathbf{x}_k^{\mathrm{T}} \mathbf{L}^{\mathrm{T}} \mathbf{R}\, \mathbf{L}\, \mathbf{u}_k)}_{J} \ < \ V(\mathbf{x}_0) = \mathbf{x}_0^{\mathrm{T}} \mathbf{P}\, \mathbf{x}_0 < \gamma ,$$

and (6.25) becomes here

$$\begin{cases} \mathbf{x}_0^{\mathrm{T}} \mathbf{P}\, \mathbf{x}_0 < \gamma \\ (\mathbf{\Phi} - \mathbf{\Gamma L})^{\mathrm{T}} \mathbf{P} (\mathbf{\Phi} - \mathbf{\Gamma L}) - \mathbf{P} + \mathbf{Q} + \mathbf{L}^{\mathrm{T}} \mathbf{R}\, \mathbf{L} < 0 \end{cases} \tag{6.30}$$

Multiplying again left and right the second inequality by $\mathbf{Y} = \mathbf{P}^{-1}$ and introducing again $\mathbf{W} = \mathbf{LY}$, this inequality writes

$$(\mathbf{Y\Phi}^{\mathrm{T}} - \mathbf{W}^{\mathrm{T}}\mathbf{\Gamma}^{\mathrm{T}})\mathbf{Y}^{-1}(\mathbf{\Phi Y} - \mathbf{\Gamma W}) - \mathbf{Y} + \mathbf{YH}^{\mathrm{T}}\mathbf{HY} + \mathbf{W}^{\mathrm{T}}\mathbf{R}\,\mathbf{W} < 0\,,$$

$$-\mathbf{Y} - \begin{pmatrix} \mathbf{Y\Phi}^{\mathrm{T}} - \mathbf{W}^{\mathrm{T}}\mathbf{\Gamma}^{\mathrm{T}} & \mathbf{YH}^{\mathrm{T}} & \mathbf{W}^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} -\mathbf{Y}^{-1} & 0 & 0 \\ 0 & -\mathbf{I} & 0 \\ 0 & 0 & -\mathbf{R} \end{pmatrix} \begin{pmatrix} \mathbf{\Phi Y} - \mathbf{\Gamma W} \\ \mathbf{HY} \\ \mathbf{W} \end{pmatrix} < 0\,,$$

and, with the Schur Lemma,

$$\begin{pmatrix} -\mathbf{Y} & \mathbf{Y\Phi}^{\mathrm{T}} - \mathbf{W}^{\mathrm{T}}\mathbf{\Gamma}^{\mathrm{T}} & \mathbf{YH}^{\mathrm{T}} & \mathbf{W}^{\mathrm{T}} \\ \mathbf{\Phi Y} - \mathbf{\Gamma W} & -\mathbf{Y} & 0 & 0 \\ \mathbf{HY} & 0 & -\mathbf{I}_q & 0 \\ \mathbf{W} & 0 & 0 & -\mathbf{R}^{-1} \end{pmatrix} < 0\ . \tag{6.31}$$

The first inequality of (6.30) being the same as in the continuous case, (6.28) holds here also.

If it is desired to minimize γ, the following optimization problem should be solved:

$$\min_{\gamma, \mathbf{Y} = \mathbf{Y}^{\mathrm{T}}, \mathbf{W}} \gamma$$
$$\text{subject to } (6.31) \text{ and } (6.28)$$

6.3.5 Pole Placement in LMI Regions

Pole assignment in convex regions in the left-half plane or in the unit disk can be expressed as LMI constraints, either applied to the Lyapunov matrix **P** involved in such a single constraint, or added to other constraints, e.g. in the case of an LQC synthesis with regional pole constraint.

We will follow here the basic developments of [ChGa96], but with the notations used in [ScGC97]. Early work in this field has been done by [ArBC93].

Kronecker product of matrices. The Kronecker product of two matrices $\mathbf{P} = \left(p_{ij}\right) \in \mathbb{R}^{m \times n}$ and $\mathbf{Q} = \left(q_{ij}\right) \in \mathbb{R}^{p \times q}$, denoted by $\otimes$, is defined as follows:

$$\mathbf{P} \otimes \mathbf{Q} = \begin{pmatrix} p_{11}\mathbf{Q} & \cdots & p_{1n}\mathbf{Q} \\ \vdots & \ddots & \vdots \\ p_{m1}\mathbf{Q} & \cdots & p_{mn}\mathbf{Q} \end{pmatrix} \in \mathbb{R}^{mp \times nq} .$$

LMI region. An LMI region is any region $\mathcal{R}$ of the complex plane that can be defined as

$$\mathcal{R} = \left\{ z \in \mathbb{C} : \mathbf{L}_{reg} + z\,\mathbf{M}_{reg} + \bar{z}\,\mathbf{M}_{reg}^{\mathrm{T}} \right\} < 0 , \tag{6.32}$$

where $\mathbf{L}_{reg} = \mathbf{L}_{reg}^{\mathrm{T}}$ and $\mathbf{M}_{reg}$ are fixed real matrices [ScGC97]. Note that $\mathcal{R}$ is a convex region in the complex plane, symmetric about the horizontal axis. Such a region can be a vertical strip, a disk, an horizontal strip, a conic sector, an ellipsoid, a parabola, and an arbitrary intersection of such regions. Details can be found in [ChGa96]. Let us give just a simple example.

Example: disk of radius *r* and center $(-q, 0)$. A point in the complex plane representing the complex number *z* belongs to this disk if and only if $\left|z+q\right| < r$, i.e. $\left|z+q\right|^2 < r^2$, which we can also write $-r^2 + (q+z)(q+\bar{z}) < 0$, or equivalently, *r* being strictly positive, $-r + (q+z)r^{-1}(q+\bar{z}) < 0$. With the Schur Lemma, this inequality can be cast in an LMI:

$$\begin{pmatrix} -r & q+z \\ q+\bar{z} & -r \end{pmatrix} < 0 .$$

By comparison with (6.32), this region is characterized by the following matrices:

$$\mathbf{L}_{reg} = \begin{pmatrix} -r & q \\ q & -r \end{pmatrix}, \quad \mathbf{M}_{reg} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} .$$

The matrices characterizing many of the above listed LMI regions are contained in the script file *yalmip_lmi_reg.m* included in the downloadable software accompanying this book. Note that the matrices characterizing a convex region defined by the intersection of several LMI regions are obtained by aligning on the diagonal the $\mathbf{L}_{reg}$ and $\mathbf{M}_{reg}$ matrices corresponding to the individual LMI regions, as results from the grouping property of a set of LMIs into one single LMI (Sect. 6.1.1).

The Lyapunov theory, which applies for the open left-half plane $\{z : z + \bar{z} < 0\}$ can be generalized to any LMI region. The eigenvalues of a given matrix **A** are all contained in the LMI region $\mathcal{R}$ if and only if some LMI involving **A** is feasible, as specified by the following theorem [ChGa96].

Theorem 6.1. The matrix **A** has all its eigenvalues in the LMI region (6.32) if and only if there exists a symmetric matrix **P** such that

$$\mathbf{L}_{reg} \otimes \mathbf{P} + \mathbf{M}_{reg} \otimes \mathbf{A}^{\mathrm{T}}\mathbf{P} + \mathbf{M}_{reg}^{\mathrm{T}} \otimes \mathbf{PA} < 0, \quad \mathbf{P} > 0 \ . \tag{6.33}$$

6.3.6 State Feedback with Regional Pole Placement

The system matrix concerned with the assignment of poles in LMI regions is here the closed-loop matrix. In case of state feedback, this matrix is $\mathbf{A} - \mathbf{BL}$ (continuous case) or $\mathbf{\Phi} - \mathbf{\Gamma L}$ (discrete case). If this matrix is substituted in (6.33), this inequality in the decision variables **L** and **P** is no longer an LMI, because of the product **LP** and its transpose. Using the same procedure as in Sect. 6.3.2.1, we multiply this inequality left and right by $\mathbf{P}^{-1}$ and introduce again $\mathbf{Y} = \mathbf{P}^{-1}$ and $\mathbf{W} = \mathbf{LY}$. The following LMI is obtained in the continuous case:

$$\mathbf{L}_{reg} \otimes \mathbf{Y} + \mathbf{M}_{reg} \otimes (\mathbf{AY} - \mathbf{BW})^{\mathrm{T}} + \mathbf{M}_{reg}^{\mathrm{T}} \otimes (\mathbf{AY} - \mathbf{BW}) < 0, \quad \mathbf{Y} > 0 \ . \tag{6.34}$$

The LMI of the discrete case is the same, with **A** and **B** replaced by $\mathbf{\Phi}$ and $\mathbf{\Gamma}$.

6.3.7 Inverse Problem of Optimal Control

The inverse problem of optimal control is the following. Given a matrix **L**, determine whether there exist two matrices, $\mathbf{Q} \geq 0$ and $\mathbf{R} > 0$, such that $(\mathbf{Q}, \mathbf{A})$ in the continuous case, respectively $(\mathbf{Q}, \mathbf{\Phi})$ in the discrete case, is detectable and $\mathbf{u} = -\mathbf{Lx}$ is the optimal control for the corresponding LQC problem.

6.3.7.1 Continuous Case

Given a continuous-time system $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu}$, $\mathbf{x}(0) = \mathbf{x}_0$, and a matrix **L**. We are seeking $\mathbf{Q} \geq 0$ and $\mathbf{R} > 0$ such that there exist $\mathbf{P} \geq 0$ and $\mathbf{P}_1 > 0$ satisfying

$$\begin{cases} (\mathbf{A}-\mathbf{BL})^{\mathrm{T}}\mathbf{P}+\mathbf{P}(\mathbf{A}-\mathbf{BL})+\mathbf{L}^{\mathrm{T}}\mathbf{R}\,\mathbf{L}+\mathbf{Q}=0 \\ \mathbf{B}^{\mathrm{T}}\mathbf{P}-\mathbf{RL}=0 \\ \mathbf{A}^{\mathrm{T}}\mathbf{P}_1+\mathbf{P}_1\mathbf{A}<\mathbf{Q} \end{cases} \tag{6.35}$$

Since **L** is given, the constraints involved here are two linear matrix *equalities* and one LMI, the decision variables being **P**, **P**$_1$, **Q** and **R**. The first constraint represents the continuous-time ARE (3.54), rewritten for linearization purposes as

$$\mathbf{PA}-\mathbf{PBR}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}+\mathbf{A}^{\mathrm{T}}\mathbf{P}-\mathbf{PBR}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}+\mathbf{PBR}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}+\mathbf{Q}=\mathbf{0}$$
$$\mathbf{P}(\mathbf{A}-\mathbf{BL})+(\mathbf{A}-\mathbf{BL})^{\mathrm{T}}\mathbf{P}+\mathbf{L}^{\mathrm{T}}\mathbf{RL}+\mathbf{Q}=\mathbf{0}\,,$$

with $\mathbf{L}=\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}$ as given by (3.55), which is reflected in the second constraint. The LMI is equivalent to the condition of $(\mathbf{Q},\mathbf{A})$ being detectable [BEFB94]. If the **L** matrix used is the one obtained by solving (3.54) and (3.55) with some arbitrary pair of weighting matrices, the matrices $\mathbf{Q}_{ipoc}$ and $\mathbf{R}_{ipoc}$ solution of (6.35) should yield the same value of **L** if the ARE problem is now solved with them.

6.3.7.2 Discrete Case

Assume given a discrete-time system $\mathbf{x}_{k+1}=\mathbf{\Phi x}_k+\mathbf{\Gamma u}_k$, $\mathbf{x}(0)=\mathbf{x}_0$, and a matrix **L**. By transposition of the previous results to the discrete situation, we seek $\mathbf{Q}\geq 0$ and $\mathbf{R}>0$ such that there exist $\mathbf{P}\geq 0$ and $\mathbf{P}_1>0$ satisfying

$$\begin{cases} (\mathbf{\Phi}-\mathbf{\Gamma L})^{\mathrm{T}}\mathbf{P}(\mathbf{\Phi}-\mathbf{\Gamma L})-\mathbf{P}+\mathbf{L}^{\mathrm{T}}\mathbf{R}\,\mathbf{L}+\mathbf{Q}=0 \\ \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}\,\mathbf{\Phi}-(\mathbf{R}+\mathbf{\Gamma}^{\mathrm{T}}\mathbf{P\Gamma})\mathbf{L}=0 \\ \mathbf{\Phi}^{\mathrm{T}}\mathbf{P}_1\mathbf{\Phi}-\mathbf{P}_1<\mathbf{Q} \end{cases} . \tag{6.36}$$

The decision variables are **P**, **P**$_1$, **Q** and **R**. The first constraint represents the discrete ARE, with $\mathbf{L}=(\mathbf{R}+\mathbf{\Gamma}^{\mathrm{T}}\mathbf{P\Gamma})^{-1}\mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}\,\mathbf{\Phi}$ (see (3.84)) reflected in the second equality. The LMI is equivalent to $(\mathbf{Q},\mathbf{\Phi})$ being detectable.

6.3.8 Extension to Uncertain Plants. Robustness

A very interesting feature of LMI methods is that they can convert elegantly a nominal design problem into a robust design. Assume, for example, that for some real plant several linear models have been developed, corresponding to different operating points, maybe also from experimental data taken at different times:

$$\dot{\mathbf{x}}_i = \mathbf{A}_i\,\mathbf{x} + \mathbf{B}_i\,\mathbf{u},\ i = 1,\ldots,n_M\ .$$

Suppose that the following robust control design is submitted: determine a matrix $\mathbf{L}$ such that the state feedback $\mathbf{u} = -\mathbf{L}\mathbf{x}$ stabilizes all these n_M models. This amounts to imposing that the closed-loop systems $\dot{\mathbf{x}}_i = (\mathbf{A}_i - \mathbf{B}_i\,\mathbf{L})\mathbf{x},\ i = 1,\ldots,n_M$ are stable. From (6.14) we can conclude that, if there exist $\mathbf{Y} = \mathbf{Y}^{\mathrm{T}} > 0$ and $\mathbf{W} = \mathbf{W}^{\mathrm{T}} > 0$ such that the following LMIs are feasible

$$\begin{cases} \mathbf{Y} > 0 \\ \mathbf{A}_i\mathbf{Y} + \mathbf{Y}\mathbf{A}_i^{\mathrm{T}} - \mathbf{B}_i\mathbf{W} - \mathbf{W}^{\mathrm{T}}\mathbf{B}_i^{\mathrm{T}} < 0,\ i = 1,\ldots,n_M \end{cases} \tag{6.37}$$

then the matrix $\mathbf{L} = \mathbf{W}\mathbf{Y}^{-1}$ solves the problem.

An identical synthesis is of course also possible for discrete-time uncertain plants, by extending this time (6.18) to n_M models of the plant.

It may be worthwhile to note that, to obtain in the above example one *single* controller it was mandatory to have only *one* Lyapunov matrix $\mathbf{Y}$ and *one* matrix $\mathbf{W}$ in the LMI (6.37), the same for all n_M partial LMIs. This represents some restriction to the use of LMIs for robust synthesis problems. Note also that the above approach, which worked well for a *state-feedback* synthesis, could not be used e.g. to treat an *output-feedback* design problem.

6.4 Software Tools to Solve LMIs

To solve under MATLAB® the LMIs presented in this chapter and used in the following exercises, a solver is needed. A powerful LMI solver is the freeware SeDuMi [Stu99]. Several versions are downloadable from the Internet sites indicated in the reference. According to their authors, Version 1.21 has been tested and confirmed to run under MATLAB® versions up to 2009a. The newly released version 1.3 supports MATLAB® versions from 2007b to 2009b. From version 1.21 on, SeDuMi runs also on 64-bit operating systems. Installation instructions are included in the downloadable archive. The exercises proposed hereafter can also be solved with CVX, an alternative freeware package for specifying and solving convex problems [GrBo08], [GrBo10], which includes SeDuMi.

A very useful free interface with MATLAB® is provided by Yalmip [Löf04]. The download link is given also in the reference. The user is invited to read carefully the installation instructions on this site. Yalmip relies on many external solvers, among which SeDuMi.

6.5 Solved Exercises

The first exercise is based on the same plant as used in previous chapters, to give the reader the possibility to compare with previous results, where comparisons are justified. The second one is more specific to the matter of the present chapter.

Exercise 6.1 Inverted Pendulum

Consider again the inverted pendulum of Exercises 1.1, 2.1 and 3.4. By using exclusively LMI methods, except in question (e), solve the following questions.

a) Determine a continuous-time state-feedback control that guarantees an α-stability for the closed loop with $\alpha = 5$. Compare with the results of Exercise 1.1, question (b).

b) Determine an LQ controller for the continuous-time model with the state weighting matrices of the first choice of Exercise 3.4, question (a), i.e., $\mathbf{Q} = 100 \times \mathbf{I}_4$, $R = 1$. Determine the rise time of the closed loop step response. Comment.

c) Calculate a state-feedback control with regional pole placement, the confinement region being defined by the intersection of the half plane to the left of $x_0 = -5$ and a disk of radius 10 centered at 0.

d) Design an LQ controller with regional pole placement, with the same weighting matrices as in question (b) and the same pole confinement specification as in question (c). Check in simulation the closed-loop step response and the corresponding control signal. Compare with the result of question (b) and comment.

e) Design an LQ controller $\mathbf{L}_{opt}$ by quadratic criterion (solution of the algebraic Riccati equation) with the same weightings as in question (b). Determine then a pair of weighting matrices $\mathbf{Q}_{ipoc}$ and $\mathbf{R}_{ipoc}$ solution of the inverse problem of optimal control, i.e. which yield this optimal control $\mathbf{L}_{opt}$. To verify that, make without exiting the program a new LQC design by solving the ARE with weightings $\mathbf{Q} = \mathbf{Q}_{ipoc}$ and $\mathbf{R} = \mathbf{R}_{ipoc}$. Compare this new $\mathbf{L}$ matrix with the $\mathbf{L}_{opt}$ calculated at the beginning of the question.

f) Repeat question (e) with the discrete-time model of the pendulum.

Solution

(a) Run *MMCE.m* with the inverted pendulum, two measurements and a continuous model. Then select LMI methods, unique component of y to be set to its reference value: 1, no integral action, alpha stabilization by state feedback, and value of alpha: 5. The state feedback controller and gain compensation matrices, obtained by solving (6.15), are:

$$\boldsymbol{\ell}^{\mathrm{T}} = \begin{pmatrix}-69.84 & 18.91 & 48.19 & 3.894\end{pmatrix}, \quad M = -69.84\,,$$

and the closed-loop poles are: $-6.204 \pm j\,2.210$, $-11.18 \pm j\,15.96$. They have all a real part less than -5, as required. In comparison with the results of Exercise 1.1, a step response at least as fast is obtained in simulation, with some overshoot, which was not the case in Exercise 1.1. This is due to the existence of complex poles here, while only real poles were obtained in Exercise 1.1. Moreover, the significantly more negative real part of the second complex pole pair, in connection with significantly larger coefficients of the feedback gain matrix, result in a much stronger control signal (-70 V initial value!). The present synthesis does not impose, indeed, any limit to the location of the poles towards high negative real parts or to high moduli.

(b) Without leaving the program, go through the same steps as in question (a) until the LMI methods menu is reached, and choose now LQ regulator design. Enter the specified weighting matrices. Solving the LMIs (6.26) and (6.28) yields the LQ controller and gain compensation matrix

$$\boldsymbol{\ell}^{\mathrm{T}} = \begin{pmatrix}-10.01 & 34.32 & 32.41 & 12.41\end{pmatrix}; \; M = -10.01\,.$$

Within an accuracy of about 0.1% the results are identical to those of Exercise 3.4, question (a), case (1). The dominant closed-loop pole is $s = -1.38$, which accounts for a rather slow rise time (time of first maximum: 4 s).

(c) Without leaving the program, go through the same steps as in the previous questions until the LMI methods menu is reached again, and choose now state-feedback design with regional pole placement, and then, for the LMI region, select successively h) Half-plane, l (left), x0 = -5, d) Disk, Abscissa q of the center: 0, radius r: 10, q) Quit. The involved LMI is here (6.34) and it is solved by the script file *yalmip_lmi_reg.m* included in the package *mmce.zip*. The following matrices are obtained for the state-feedback control and the gain compensation:

$$\boldsymbol{\ell}^{\mathrm{T}} = \begin{pmatrix}-11.91 & 4.951 & 11.29 & 1.067\end{pmatrix}; \; M = -11.91\,.$$

If we compare with question (a), the coefficients of the state feedback gain matrix are now much closer to those obtained by pole placement in Exercise 1.1 or by LQC synthesis, and the control signal stays within reasonable bounds (-12 V initial value). The closed loop poles are now:

$$-5.575 \pm j\,1.788\,;\ -5.776 \pm j\,6.965\,,$$

their respective moduli obtained by abs(Poles) being: 5.854, 5.854, 9.048, and 9.048, which indicates that the specifications are satisfied.

(d) Still without leaving the program, go through the menus until reaching the LMI methods menu. Choose now LQ regulator design with regional pole placement, with the same choices as previously for the weighting matrices and the given constraints for the LMI region imposed to the closed-loop poles. By solving (6.26) and (6.28) again, with addition by yalmip_lmi_reg of the closed-loop pole confinement constraint (6.34), the program determines the following solution:

$$\boldsymbol{\ell}^{\mathrm{T}} = \begin{pmatrix} -11.34 & 4.74 & 10.77 & 1.02 \end{pmatrix}, \quad M = -11.34\,.$$

The closed-loop poles are $-5.431 \pm j\,1.834$, $-5.528 \pm j\,7.125$, which respects fully the regional placement constraint. In comparison with question (b), the dominant pole pair (closest to the origin), has a modulus of 5.73, resulting now in a much faster rise time (time of first maximum: 0.7 s).

(e) In order to dispose of an LQ controller for this question, retrieve first the state-feedback matrix obtained by following the steps of Exercise 3.4, question (a), case (1). Then, without leaving the program, proceed again to the LMI methods menu and choose inverse problem of optimal control (ipoc). Answer yes to the question do you want to use the last calculated L matrix. The following weighting matrices are obtained:

$$\mathbf{Q}_{ipoc} = \begin{pmatrix} 0.6561 & -0.1390 & -0.0224 & 0.0206 \\ -0.1390 & 1.7460 & 0.4293 & 0.5089 \\ -0.0224 & 0.4293 & 1.0355 & 0.3910 \\ 0.0206 & 0.5089 & 0.3910 & 0.7110 \end{pmatrix}, \quad R_{ipoc} = 0.0066\,.$$

If they are now taken as weighting matrices in a new state feedback synthesis, by quadratic criterion — answering this time “n” to the suggested choice of a diagonal form, both for **Q** and **R**, and typing simply Q_ipoc and R_ipoc at the next prompts – the state feedback gain $\boldsymbol{\ell}^{\mathrm{T}} = \begin{pmatrix} -10.00 & 34.28 & 32.37 & 12.40 \end{pmatrix}$ results, with the following closed loop poles: $-845.3,\ -1.38,\ -0.92 \pm j1.09$. These results are identical to those obtained in Exercise 3.4.

(f) The user being now familiar with the use of *MMCE.m*, we leave it to him to treat this question in discrete-time. For the sake of verification, we give here just the state-feedback matrix obtained with an LQC design reusing the weighting matrices $\mathbf{Q}_{ipoc}$ and $\mathbf{R}_{ipoc}$: $\ell^{\mathrm{T}} = \begin{pmatrix} -0.3857 & 1.6253 & 1.5713 & 0.5187 \end{pmatrix}$.

Exercise 6.2 Uncontrollable and Unobservable Example

The plant is here the following third order academic model, which is open loop unstable, not completely controllable and not completely observable:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -4 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{C}_m = \begin{pmatrix} 0.5 & 0.5 & 0 \end{pmatrix}.$$

It is nevertheless stabilizable and detectable in this initial form. The program *MMCE.m* gives the user the possibility to change it to unstabilizable or to undetectable form, or to both simultaneously

a) This plant being uncontrollable, check its stabilizability and determine a stabilizing state-feedback controller by LMI method. The plant being fed back directly by this state-feedback controller, observe by simulation the step response of the three state variables. Comment. Apply now successively a load and a measurement disturbance, and comment your observations.

b) Without quitting the program, check the detectability of this plant in its initial form and determine an observer by LMI method. The plant being now controlled by means of this observer and the controller determined in the previous question, simulate its step response again and comment. To study the behavior of the observer, you can let it operate in free wheel, the loop being closed from the state.

c) Restart the program and change the plant model to “unstabilizable”. Repeat the manipulations of questions (a) and (b).

d) Restart the program and change the plant model to “undetectable”. Repeat the manipulations of questions (a) and (b).

Solution:

(a) Run *MMCE.m* by choosing the plant academic example, then the plant type 0) stabilizable and detectable, continuous model, LMI methods, without integral action, and, among LMI methods, choose stabilizability and stabilization by state feedback. The program solves then (6.14) and yields the state-feedback controller and gain compensation matrix

$$\boldsymbol{\ell}^{\mathrm{T}} = (1.7111 \quad 0 \quad -1.1829), \quad M = 2.0136\,.$$

The fact that the second component of $\boldsymbol{\ell}^{\mathrm{T}}$ vanishes is not surprising: x_2 is indeed the uncontrolled state variable as shown by the expression of **B** in the diagonal representation of the plant. The simulated step response is stable. A load disturbance applied evenly to all three states does not destabilize the closed loop, as expected since the uncontrolled mode is itself stable ($\lambda_2 = -2$). The disturbance is of course not rejected, since no means have been taken for that.

(b) Without leaving the program, go through the same steps as in question (a) until the LMI methods menu is reached, and choose now detectability and detection by an observer. The unobservable mode being stable ($\lambda_3 = -4$), an observer is obtained by solving (6.20) with the following gain matrix:

$$\mathbf{g}^{\mathrm{T}} = \left(4.1321 \quad -0.4147 \quad 0\right).$$

The measurement does not influence the estimation of x_3, as expected, this state variable being the unobserved one from the output as indicated by the expression of $\mathbf{C}_m$ in the diagonal representation of the plant. With the observer in the loop, the simulated closed-loop step response is stable, as in the previous question. However, when the observer is operated in free wheel, it does not estimate correctly x_3 when a load disturbance is applied, due to the lack of information given on that variable by **g**.

(c) Though the LMI solver mentions that it detected no problem, the algorithm did not converge towards a usable solution, leading to the conclusion that the plant is no longer stabilizable, as expected, since now the uncontrollable state variable, x_2 corresponds to an unstable mode ($\lambda_2 = 2$).

(d) The algorithm does not converge either: the plant is now undetectable, since the mode corresponding to x_3 is now unstable ($\lambda_3 = 4$).

A State Space Representations of Systems

The formulae and theorems recalled in this Appendix are given mostly without demonstration. For more details the reader is invited to consult one of the manuals listed in the References at the end of this book, such as [Oga02], [Ost04], [CAJZ01], for continuous-time systems or [AsWi97] for discrete-time systems.

A.1 State-space Representations of Linear Time-invariant (LTI) Systems

A.1.1 Canonical State Representations of SISO Systems

Among the infinite number of possible state space representations of a linear, time-invariant system, there are three *canonical* forms. They are given in Table A.1 for a SISO system in connection with the parameters of its transfer function. The continuous-time and discrete-time cases are presented in parallel in this table, as well as in the other tables of this Appendix.

The two systems are assumed of order n (number of state variables required to describe them), which corresponds to a denominator of their transfer functions of degree n. For the sake of generality, the transfer functions are entered into the table with the same numerator and denominator degrees, which results in a coefficient $b_n \neq 0$. In most practical cases, however, due to physical reasons (finite bandwidth), the numerator degree is less than the denominator degree by at least one unit, which makes then $b_n = 0$.

We will assume furthermore that the rational fractions have been simplified by the coefficient a_n of the denominator's highest degree term, resulting in the assumption $a_n = 1$.

E. Ostertag, *Mono- and Multivariable Control and Estimation,*
DOI 10.1007/978-3-642-13734-1_7, © Springer-Verlag Berlin Heidelberg 2011

Table A.1 The three canonical state representations of a SISO system.

Continuous case	Discrete case
Transfer function (assumption: $a_n = 1$):	
$Y(s) = G(s)U(s);$ $G(s) = \dfrac{b_{n-1}s^{n-1} + \cdots + b_1 s + b_0}{s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0} + b_n$	$Y(z) = G(z)U(z);$ $G(z) = \dfrac{b_{n-1}z^{n-1} + \cdots + b_1 z + b_0}{z^n + a_{n-1}z^{n-1} + \cdots + a_1 z + a_0} + b_n$
Controllability canonical form (assumption: $a_n = 1$):	
$\dot{\mathbf{x}} = \underbrace{\begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 \\ -a_0 & -a_1 & \cdots & -a_{n-1} \end{pmatrix}}_{\mathbf{A}_{\mathcal{C}}} \mathbf{x} + \underbrace{\begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}}_{\mathbf{b}_{\mathcal{C}}} u$ $y = \underbrace{\begin{pmatrix} b_0 & b_1 & \cdots & b_{n-1} \end{pmatrix}}_{\mathbf{c}_{\mathcal{C}}^{\mathrm{T}}} \mathbf{x} + \underbrace{b_n}_{\mathbf{d}_{\mathcal{C}}} u$ (A.1)	$\mathbf{x}_{k+1} = \underbrace{\begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 \\ -a_0 & -a_1 & \cdots & -a_{n-1} \end{pmatrix}}_{\mathbf{\Phi}_{\mathcal{C}}} \mathbf{x}_k + \underbrace{\begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}}_{\boldsymbol{\gamma}_{\mathcal{C}}} u_k$ $y_k = \underbrace{\begin{pmatrix} b_0 & b_1 & \cdots & b_{n-1} \end{pmatrix}}_{\mathbf{c}_{\mathcal{C}}^{\mathrm{T}}} \mathbf{x}_k + \underbrace{b_n}_{\mathbf{d}_{\mathcal{C}}} u_k$ (A.2)
Observability canonical form (assumption: $a_n = 1$):	
$\dot{\mathbf{x}} = \underbrace{\begin{pmatrix} 0 & \cdots & \cdots & -a_0 \\ 1 & \ddots & & -a_1 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 1 & -a_{n-1} \end{pmatrix}}_{\mathbf{A}_{\mathcal{O}}} \mathbf{x} + \underbrace{\begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{n-1} \end{pmatrix}}_{\mathbf{b}_{\mathcal{O}}} u$ $y = \underbrace{\begin{pmatrix} 0 & \cdots & 0 & 1 \end{pmatrix}}_{\mathbf{c}_{\mathcal{O}}^{\mathrm{T}}} \mathbf{x} + \underbrace{b_n}_{\mathbf{d}_{\mathcal{O}}} u$ (A.3)	$\mathbf{x}_{k+1} = \underbrace{\begin{pmatrix} 0 & \cdots & \cdots & -a_0 \\ 1 & \ddots & & -a_1 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 1 & -a_{n-1} \end{pmatrix}}_{\mathbf{\Phi}_{\mathcal{O}}} \mathbf{x}_k + \underbrace{\begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{n-1} \end{pmatrix}}_{\boldsymbol{\gamma}_{\mathcal{O}}} u_k$ $y_k = \underbrace{\begin{pmatrix} 0 & \cdots & 0 & 1 \end{pmatrix}}_{\mathbf{c}_{\mathcal{O}}^{\mathrm{T}}} \mathbf{x}_k + \underbrace{b_n}_{\mathbf{d}_{\mathcal{O}}} u_k$ (A.4)
Diagonal form (assumption: poles of $G(s)$ or $G(z)$ all distinct, real or complex):	
$\dot{\mathbf{x}} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & \lambda_n \end{pmatrix} \mathbf{x} + \begin{pmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{pmatrix} u$ $y = \begin{pmatrix} r_1 & r_2 & \cdots & r_n \end{pmatrix} \mathbf{x} + r_0 u$ (A.5)	$\mathbf{x}_{k+1} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & \lambda_n \end{pmatrix} \mathbf{x}_k + \begin{pmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{pmatrix} u_k$ $y = \begin{pmatrix} r_1 & r_2 & \cdots & r_n \end{pmatrix} \mathbf{x}_k + r_0 u_k$ (A.6)

A.1.2 General Form of the State Equations (MIMO Systems)

The general state representation of a system of order n, having p inputs and q outputs, is given in Table A.2. For the sake of generality, the assumption $\mathbf{D} \neq 0$ has been made in this table, which is the same assumption as $b_n \neq 0$ in the SISO case. In most practical situations, however, this direct feedthrough term from input to output will vanish, i.e. $\mathbf{D}=0$.

Table A.2 General state space representation of a MIMO system.

Continuous case		Discrete case	
$\begin{cases}\dot{\mathbf{x}}(t)=\mathbf{A}\mathbf{x}(t)+\mathbf{B}\mathbf{u}(t)\\ \mathbf{y}(t)=\mathbf{C}\mathbf{x}(t)+\mathbf{D}\mathbf{u}(t)\end{cases}$	(A.7)	$\begin{cases}\mathbf{x}_{k+1}=\mathbf{\Phi}\mathbf{x}_k+\mathbf{\Gamma}\mathbf{u}_k\\ \mathbf{y}_k=\mathbf{C}\mathbf{x}_k+\mathbf{D}\mathbf{u}_k\end{cases}$	(A.8)

with:

$$\text{length}(\mathbf{x})=n,\ \text{length}(\mathbf{u})=p,\ \text{length}(\mathbf{y})=q$$

$$\text{size}(\mathbf{A}\text{ or }\mathbf{\Phi})=n\times n,\ \text{size}(\mathbf{B}\text{ or }\mathbf{\Gamma})=n\times p,\ \text{size}(\mathbf{C})=q\times n,\ \text{size}(\mathbf{D})=q\times p$$

and:

$$p\le n;\ \ q\le n;\ \ \text{rank}(\mathbf{B}\text{ or }\mathbf{\Gamma})=p;\ \ \text{rank}(\mathbf{C})=q$$

The sampling with zero-order hold at a period T_s of a continuous-time system yields the following matrices replacing **A** and **B**, whereas **C** and **D** remain unchanged:

$$\begin{aligned}\mathbf{\Phi}&=e^{\mathbf{A}T_s}\\ \mathbf{\Gamma}&=\int_0^{T_s}e^{\mathbf{A}\tau}d\tau\cdot\mathbf{B}=(\mathbf{\Phi}-\mathbf{I})\mathbf{A}^{-1}\mathbf{B},\ \text{if }\mathbf{A}^{-1}\text{ exists}\end{aligned}\tag{A.9}$$

Proof. See Sect. A.5.

A.2 Controllability of Linear Time-invariant Systems

A.2.1 Definition

A linear, time-invariant system represented by (A.7), respectively by (A.8), where **A**, **B**, **C** and **D**, respectively $\mathbf{\Phi}$, $\mathbf{\Gamma}$, **C** and **D**, are constant matrices, is said to be (completely) controllable, if its state **x** can be transferred, by an appropriate choice of the input vector **u**, in a finite amount of time, from an arbitrary initial state $\mathbf{x}(t_0)$ to an arbitrary final state $\mathbf{x}(t_f)$.

A.2.2 Controllability Criterion

A mathematical controllability criterion is given in Table A.3. A practical criterion will follow it.

Table A.3 Mathematical controllability criterion.

Continuous case	Discrete case
The necessary and sufficient condition for the system	
$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$ ($\mathbf{A}, \mathbf{B}$: constant)	$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}\mathbf{u}_k$ ($\mathbf{\Phi}, \mathbf{\Gamma}$: constant)
to be controllable is that the $(n \times np)$ *controllability matrix*	
$\mathbf{Q}_C = \begin{pmatrix} \mathbf{B} & \mathbf{A}\mathbf{B} & \cdots & \mathbf{A}^{n-1}\mathbf{B} \end{pmatrix}$ (A.10)	$\mathbf{Q}_C = \begin{pmatrix} \mathbf{\Gamma} & \mathbf{\Phi}\mathbf{\Gamma} & \cdots & \mathbf{\Phi}^{n-1}\mathbf{\Gamma} \end{pmatrix}$ (A.11)
is of rank *n*, thus has *n* linearly independent columns. N.B.: in case of a *single input* system, the matrix	
$\mathbf{Q}_C = \begin{pmatrix} \mathbf{b} & \mathbf{A}\mathbf{b} & \cdots & \mathbf{A}^{n-1}\mathbf{b} \end{pmatrix}$ (A.12)	$\mathbf{Q}_C = \begin{pmatrix} \boldsymbol{\gamma} & \mathbf{\Phi}\boldsymbol{\gamma} & \cdots & \mathbf{\Phi}^{n-1}\boldsymbol{\gamma} \end{pmatrix}$ (A.13)
must be regular ($\Leftrightarrow \det \mathbf{Q}_C \neq 0$).	

Comment about the controllability test: practical controllability. Though for academic examples the computation of $\mathrm{rank}(\mathbf{Q}_C)$ is good enough, this may not be the case for real industrial plants, bearing some uncertainty in the model matrices. In the case $\mathbf{Q}_C$ has a large condition number (see Sect. B.5, Appendix B), a slight change in one of the model parameters might change the conclusion from controllable to uncontrollable. It is then safer to check this situation by applying a singular value decomposition to $\mathbf{Q}_C$ (see Sect. B.5), and rounding to zero too small singular values, thus obtaining a reduced, but more realistic, rank of this matrix. An interesting example is given in [AlSa04], Example A.8.

A.2.3 Stabilizability Concept

The word *completely* has been put between parentheses in the previous definition because it is often omitted, but implicitly present notwithstanding any other indication. Assume now that a system has a controllability matrix with a rank $n_C < n$. It is thus not *completely* controllable. By means of an appropriate change of variable $\mathbf{x} = \mathbf{T}\mathbf{z}$, (see Sect. A.7.1), it is possible to put it in the following form ([Lar93], [ZhDG96], [Duc01]):

$$\begin{cases} \begin{pmatrix} \dot{\mathbf{z}}_C \\ \dot{\mathbf{z}}_{\bar{C}} \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{A}}_C & \hat{\mathbf{A}}_{12} \\ 0 & \hat{\mathbf{A}}_{\bar{C}} \end{pmatrix} \begin{pmatrix} \mathbf{z}_C \\ \mathbf{z}_{\bar{C}} \end{pmatrix} + \begin{pmatrix} \hat{\mathbf{B}}_C \\ 0 \end{pmatrix} \mathbf{u} \\ \mathbf{y} = \begin{pmatrix} \hat{\mathbf{C}}_C & \hat{\mathbf{C}}_{\bar{C}} \end{pmatrix} \begin{pmatrix} \mathbf{z}_C \\ \mathbf{z}_{\bar{C}} \end{pmatrix} \end{cases} \quad \text{or} \quad \begin{cases} \begin{pmatrix} \mathbf{z}_{C,k+1} \\ \mathbf{z}_{\bar{C},k+1} \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{\Phi}}_C & \hat{\mathbf{\Phi}}_{12} \\ 0 & \hat{\mathbf{\Phi}}_{\bar{C}} \end{pmatrix} \begin{pmatrix} \mathbf{z}_{C,k} \\ \mathbf{z}_{\bar{C},k} \end{pmatrix} + \begin{pmatrix} \hat{\mathbf{\Gamma}}_C \\ 0 \end{pmatrix} \mathbf{u}_k \\ \mathbf{y}_k = \begin{pmatrix} \hat{\mathbf{C}}_C & \hat{\mathbf{C}}_{\bar{C}} \end{pmatrix} \begin{pmatrix} \mathbf{z}_{C,k} \\ \mathbf{z}_{\bar{C},k} \end{pmatrix} \end{cases},$$

where size($\hat{\mathbf{A}}_C$ or $\hat{\mathbf{\Phi}}_C) = n_C \times n_C$, and where $\mathbf{z}_C$ groups together the controllable state variables: the pair $(\hat{\mathbf{A}}_C, \hat{\mathbf{B}}_C)$, or $(\hat{\mathbf{\Phi}}_C, \hat{\mathbf{\Gamma}}_C)$, is (completely) controllable. The state variables $\mathbf{z}_{\bar{C}}$ are not influenced by the control **u**: these variables are thus not controllable. The above equations are the *controllability canonical decomposition.*

A system is said to be *stabilizable* if its uncontrollable state variables have a stable dynamics, or, equivalently, if all its unstable modes are controllable.

Determination of this canonical decomposition. For a system of order *n* for which $\text{rank}(\mathbf{Q}_C) = n_C < n$, the n_C independent columns of $\mathbf{Q}_C$ form a basis of the controllable subspace. The uncontrollable subspace will then be spanned by the $n - n_C$ vectors orthogonal to the columns of $\mathbf{Q}_C$, i.e. by a basis of the null space of $\mathbf{Q}_C^{\mathrm{T}}$ obtained by solving $\mathbf{Q}_C^{\mathrm{T}}\mathbf{x} = 0$ (see Appendix B, Sect. B.4, and particularly Property B.10). This can be done by the MATLAB® statement null(Qc’) or, in case $\mathbf{Q}_C$ has a large condition number, by singular value decomposition, svd(Qc’) (see Sect. B.5).

To establish the similitude transformation **T** to the above canonical representation it is advantageous to make use of the orthogonal complementarity property of both subspaces (Sect. B.4), as suggested in [AlSa04]. One would then compute, successively, in the MATLAB® environment:

- basis of the uncontrollable subspace: T1 = null(Qc’);
- basis of the controllable subspace: T2 = null(T1’).

The transformation which yields the controllability canonical decomposition is then $\mathbf{x} = \mathbf{T}\mathbf{z}$ where $\mathbf{T} = \begin{pmatrix}\mathbf{T}_2 & \mathbf{T}_1\end{pmatrix}$.

A.3 Observability of Linear Time-invariant Systems

A.3.1 Definition

A linear time-invariant system represented by (A.7), respectively by (A.8), where **A**, **B**, **C** and **D**, respectively **Φ**, **Γ**, **C** and **D**, are constant matrices, is said to be (completely) observable, if it is possible, for a given $\mathbf{u}(t)$, to deduce the initial state $\mathbf{x}(t_0)$, whatever it is, from the measurement of the output $\mathbf{y}(t)$ during a finite time interval $t_0 \le t \le t_a$.

A.3.2 Observability Criterion

A mathematical observability criterion is given in Table A.4. For practical situations, the same comment applies as above for the controllability, by duality: the singular value decomposition of the observability matrix is then recommended.

Table A.4 Mathematical observability criterion.

Continuous case	Discrete case
The necessary and sufficient condition for the system	
$\begin{cases}\dot{\mathbf{x}} = \mathbf{A}\,\mathbf{x} + \mathbf{B}\,\mathbf{u}\\ \mathbf{y} = \mathbf{C}\,\mathbf{x} + \mathbf{D}\,\mathbf{u}\end{cases}$, (**A**, **B**, **C**, **D** constant)	$\begin{cases}\mathbf{x}_{k+1} = \mathbf{\Phi}\,\mathbf{x}_k + \mathbf{\Gamma}\,\mathbf{u}_k\\ \mathbf{y}_k \;\;\; = \mathbf{C}\,\mathbf{x}_k + \mathbf{D}\,\mathbf{u}_k\end{cases}$, ($\mathbf{\Phi},\mathbf{\Gamma},\mathbf{C},\mathbf{D}$ constant)
to be observable is that the $(nq\times n)$ *observability matrix*	
$\mathbf{Q}_{\mathcal{O}} = \begin{pmatrix}\mathbf{C}\\ \mathbf{C}\,\mathbf{A}\\ \vdots\\ \mathbf{C}\,\mathbf{A}^{n-1}\end{pmatrix}$ (A.14)	$\mathbf{Q}_{\mathcal{O}} = \begin{pmatrix}\mathbf{C}\\ \mathbf{C}\mathbf{\Phi}\\ \vdots\\ \mathbf{C}\mathbf{\Phi}^{n-1}\end{pmatrix}$ (A.15)
is of rank *n*, thus has *n* linearly independent rows.	
N.B.: in case of a *single output* system, the matrix	
$\mathbf{Q}_{\mathcal{O}} = \begin{pmatrix}\mathbf{c}^{\mathrm{T}}\\ \mathbf{c}^{\mathrm{T}}\mathbf{A}\\ \vdots\\ \mathbf{c}^{\mathrm{T}}\mathbf{A}^{n-1}\end{pmatrix}$ (A.16)	$\mathbf{Q}_{\mathcal{O}} = \begin{pmatrix}\mathbf{c}^{\mathrm{T}}\\ \mathbf{c}^{\mathrm{T}}\mathbf{\Phi}\\ \vdots\\ \mathbf{c}^{\mathrm{T}}\mathbf{\Phi}^{n-1}\end{pmatrix}$ (A.17)
must be regular ($\Leftrightarrow$ det $\mathbf{Q}_{\mathcal{O}} \neq 0$).	

A.3.3 Detectability Concept

Consider a not completely observable system, having an observability matrix of rank $n_{\mathcal{O}} < n$. By means of an appropriate change of variables $\mathbf{x} = \mathbf{T}\,\mathbf{z}$, it is possible to put it in the following form ([Lar93], [ZhDG96], [Duc01]):

$$\begin{cases}\begin{pmatrix}\dot{\mathbf{z}}_{\mathcal{O}}\\ \dot{\mathbf{z}}_{\bar{\mathcal{O}}}\end{pmatrix} = \begin{pmatrix}\hat{\mathbf{A}}_{\mathcal{O}} & 0\\ \hat{\mathbf{A}}_{21} & \hat{\mathbf{A}}_{\bar{\mathcal{O}}}\end{pmatrix}\begin{pmatrix}\mathbf{z}_{\mathcal{O}}\\ \mathbf{z}_{\bar{\mathcal{O}}}\end{pmatrix} + \begin{pmatrix}\hat{\mathbf{B}}_{\mathcal{O}}\\ \hat{\mathbf{B}}_{\bar{\mathcal{O}}}\end{pmatrix}\mathbf{u}\\ \mathbf{y} = \begin{pmatrix}\hat{\mathbf{C}}_{\mathcal{O}} & 0\end{pmatrix}\begin{pmatrix}\mathbf{z}_{\mathcal{O}}\\ \mathbf{z}_{\bar{\mathcal{O}}}\end{pmatrix}\end{cases} \text{ or } \begin{cases}\begin{pmatrix}\mathbf{z}_{\mathcal{O},k+1}\\ \mathbf{z}_{\bar{\mathcal{O}},k+1}\end{pmatrix} = \begin{pmatrix}\hat{\mathbf{\Phi}}_{\mathcal{O}} & 0\\ \hat{\mathbf{\Phi}}_{21} & \hat{\mathbf{\Phi}}_{\bar{\mathcal{O}}}\end{pmatrix}\begin{pmatrix}\mathbf{z}_{\mathcal{O},k}\\ \mathbf{z}_{\bar{\mathcal{O}},k}\end{pmatrix} + \begin{pmatrix}\hat{\mathbf{\Gamma}}_{\mathcal{O}}\\ \hat{\mathbf{\Gamma}}_{\bar{\mathcal{O}}}\end{pmatrix}\mathbf{u}_k\\ \mathbf{y}_k = \begin{pmatrix}\hat{\mathbf{C}}_{\mathcal{O}} & 0\end{pmatrix}\begin{pmatrix}\mathbf{z}_{\mathcal{O},k}\\ \mathbf{z}_{\bar{\mathcal{O}},k}\end{pmatrix}\end{cases},$$

where size($\hat{\mathbf{A}}_{\mathcal{O}}$ or $\hat{\mathbf{\Phi}}_{\mathcal{O}}) = n_{\mathcal{O}} \times n_{\mathcal{O}}$, and where $\mathbf{z}_{\mathcal{O}}$ groups together the observable state variables: the pair $(\hat{\mathbf{A}}_{\mathcal{O}}, \hat{\mathbf{C}}_{\mathcal{O}})$, or $(\hat{\mathbf{\Phi}}_{\mathcal{O}}, \hat{\mathbf{C}}_{\mathcal{O}})$, is (completely) observable.

The output $\mathbf{y}$ depends only on the state variables $\mathbf{z}_{\mathcal{O}}$, which themselves are not influenced by the variables $\mathbf{z}_{\bar{\mathcal{O}}}$: these last state variables are therefore not observable. The above form is the *observability canonical decomposition*.

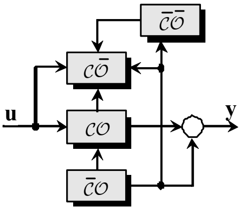
A system is said to be *detectable* if its unobservable state variables have a stable dynamics, or, equivalently, if all its unstable modes are observable.

Determination of this canonical decomposition. The approach is the same, by duality, as for the controllability situation. Let us just mention here that the unobservable subspace is spanned by a basis of the null space of $\mathbf{Q}_{\mathcal{O}}$, i.e. by the solutions of $\mathbf{Q}_{\mathcal{O}}\mathbf{x} = 0$. Such a basis can be obtained by the MATLAB® statements null(Qo) or svd(Qo) (see Sect. B.5).The observable subspace, which is orthogonal to it, is spanned by the rows of $\mathbf{Q}_{\mathcal{O}}$, i.e. by the columns of $\mathbf{Q}_{\mathcal{O}}^{\mathrm{T}}$.

A.4 Kalman Canonical Decomposition

This decomposition consists in splitting the state vector into four parts ([ZhDG96], [AsWi97]) as sketched on the right, reflecting the four possible combinations, with the following subscripts:

controllable and observable: co;
controllable but unobservable: $c\bar{o}$;
uncontrollable but observable: $\bar{c}o$;
uncontrollable and unobservable: $\bar{c}\bar{o}$.



In case of a continuous time LTI system, there exists then a coordinate transformation $\mathbf{z} = \mathbf{T}\mathbf{x}$ such that

$$\begin{cases} \begin{pmatrix} \dot{\mathbf{z}}_{co} \\ \dot{\mathbf{z}}_{c\bar{o}} \\ \dot{\mathbf{z}}_{\bar{c}o} \\ \dot{\mathbf{z}}_{\bar{c}\bar{o}} \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{A}}_{co} & 0 & \hat{\mathbf{A}}_{13} & 0 \\ \hat{\mathbf{A}}_{21} & \hat{\mathbf{A}}_{c\bar{o}} & \hat{\mathbf{A}}_{23} & \hat{\mathbf{A}}_{24} \\ 0 & 0 & \hat{\mathbf{A}}_{\bar{c}o} & 0 \\ 0 & 0 & \hat{\mathbf{A}}_{43} & \hat{\mathbf{A}}_{\bar{c}\bar{o}} \end{pmatrix} \begin{pmatrix} \mathbf{z}_{co} \\ \mathbf{z}_{c\bar{o}} \\ \mathbf{z}_{\bar{c}o} \\ \mathbf{z}_{\bar{c}\bar{o}} \end{pmatrix} + \begin{pmatrix} \hat{\mathbf{B}}_{co} \\ \hat{\mathbf{B}}_{c\bar{o}} \\ 0 \\ 0 \end{pmatrix} \mathbf{u} \\ \mathbf{y} = \begin{pmatrix} \hat{\mathbf{C}}_{co} & 0 & \hat{\mathbf{C}}_{\bar{c}o} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{z}_{co} \\ \mathbf{z}_{c\bar{o}} \\ \mathbf{z}_{\bar{c}o} \\ \mathbf{z}_{\bar{c}\bar{o}} \end{pmatrix} \end{cases}$$

Of course, a similar approach exists for discrete time systems.

The transfer matrix from **u** to **y** is given by

$$\mathbf{G}(s) = \hat{\mathbf{C}}_{CO}(s\mathbf{I} - \hat{\mathbf{A}}_{CO})^{-1}\hat{\mathbf{B}}_{CO} \,.$$

It is important to remember that, even though the transfer matrix of a dynamical system involves only its controllable and observable part as shown by this equation, the internal behaviour, e.g. the response to nonzero initial conditions or to disturbances, is quite different.

A triplet **A**, **B**, **C**, which is both controllable and observable and which satisfies $\mathbf{G}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$, is called *a minimal realization* of $\mathbf{G}(s)$.

An elegant way of determining the similitude transformation which yields the above decomposition has been given in [AlSa04] in Example 3.12. It relies again on the computation of the null spaces of the controllability and observability matrices by the MATLAB null command, or by their singular value decomposition (svd) if there is doubt about the plant's *practical* controllability and/or observability. The successive bases are obtained as follows:

- uncontrollable and unobservable part: Nncno=null([Qc';Qo]);
- uncontrollable but observable part: Nnco=null([Qc';Nncno']);
- controllable but unobservable part: Ncno=null([Qo;Nncno']);
- controllable and observable part: Nco=null([Nncno';Nnco';Ncno']).

The similitude transformation is then: T=[Nco Ncno Nnco Nncno], and the matrices of the canonical decomposition are obtained by (A.30).

A.5 Solving the State Equations in Time Domain

The total response to an initial state $\mathbf{x}_0$ at $t_0 = 0$ and an input $\mathbf{u}(t)$ or $\mathbf{u}_k$ is:

continuous case:
$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}_0 + \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)d\tau \,. \tag{A.18}$$

discrete case:
$$\mathbf{x}_k = \mathbf{\Phi}^k\,\mathbf{x}_0 + \sum_{i=0}^{k-1}\mathbf{\Phi}^{k-1-i}\,\mathbf{\Gamma}\,\mathbf{u}_i \,. \tag{A.19}$$

The proof of (A.18) is similar to solving a scalar differential equation, by left multiplying (A.7) by $e^{-\mathbf{A}t}$, [Ost04]. A *matrix exponential* is defined by:

$$e^{\mathbf{A}} = \sum_{k=0}^{\infty}\frac{\mathbf{A}^k}{k!} = \mathbf{I} + \frac{\mathbf{A}}{1!} + \frac{\mathbf{A}^2}{2!} + \cdots \;, \tag{A.20}$$

and has the following property in a change of basis (see Sect. A.7):

$$e^{\mathbf{T}^{-1}\mathbf{A}\mathbf{T}} = \mathbf{T}^{-1}e^{\mathbf{A}}\,\mathbf{T}, \quad \forall \mathbf{T} \text{ invertible}\,, \tag{A.21}$$

the proof of which is trivial by replacing **A** in (A.20) by $\mathbf{T}^{-1}\mathbf{A}\mathbf{T}$. (A.19) is established by repetitive application of (A.8) from the initial state $\mathbf{x}_0$ until time step *k*.

Application. Proof of (A.9).
By applying (A.18) to the interval $kT_s \le t < (k+1)T_s$, thus with the replacement $\mathbf{x}_0 = \mathbf{x}_k$, we obtain:

$$\mathbf{x}_{k+1} = e^{\mathbf{A}T_s}\;\mathbf{x}_k + \int_{kT_s}^{(k+1)T_s} e^{\mathbf{A}\left[(k+1)T_s-\tau\right]}\;\mathbf{B}\,\mathbf{u}(\tau)\,d\tau\,. \tag{A.22}$$

Introduce now

$$\mathbf{\Phi} = e^{\mathbf{A}T_s}\,, \tag{A.23}$$

and assume that the input $\mathbf{u}(t) = \mathbf{u}_k$ is constant for $kT_s \le t < (k+1)T_s$. By factorizing the constant terms out of the integral, (A.22) can be written

$$\mathbf{x}_{k+1} = \mathbf{\Phi}\,\mathbf{x}_k + \int_{kT_s}^{(k+1)T_s} e^{\mathbf{A}\left[(k+1)T_s-\tau\right]}\;d\tau\cdot\mathbf{B}\,\mathbf{u}_k\,. \tag{A.24}$$

The change of variable $\tau' = (k+1)T_s - \tau$ in the integral yields

$$\int_{kT_s}^{(k+1)T_s} e^{\mathbf{A}\left[(k+1)T_s-\tau\right]}\;d\tau = -\int_{T_s}^{0} e^{\mathbf{A}\tau'}d\tau' = \int_{0}^{T_s} e^{\mathbf{A}\tau'}d\tau'\,. \tag{A.25}$$

If **A** is invertible, this integral is evaluated as follows:

$$\int_{0}^{T_s} e^{\mathbf{A}\tau}d\tau = \mathbf{A}^{-1}\,e^{\mathbf{A}\tau}\Big|_{0}^{T_s} = \mathbf{A}^{-1}(e^{\mathbf{A}T_s} - \mathbf{I}) = (\mathbf{\Phi} - \mathbf{I})\,\mathbf{A}^{-1}\,. \tag{A.26}$$

(A.23) to (A.26) prove (A.9).

Transition matrix (free response). $\mathbf{\Phi}(t) = e^{\mathbf{A}t}$, or $e^{\mathbf{A}(t-t_0)}$ if $t_0 \ne 0$, respectively $\mathbf{\Phi}^k$, is called *transition matrix*, since it produces the free transition of the system state from $\mathbf{x}_0$ at initial time t_0 to $\mathbf{x}(t)$ at time *t*.
For *time varying* systems, where the value of the initial time matters, one defines a transition matrix $\mathbf{\Phi}(t,t_0)$, such that $\mathbf{x}(t) = \mathbf{\Phi}(t,t_0)\mathbf{x}(t_0)$. An evident property is: $\mathbf{\Phi}(t_2,t_1)\mathbf{\Phi}(t_1,t_0) = \mathbf{\Phi}(t_2,t_0)$, and another one is that $\mathbf{\Phi}(t,t_0)$ satisfies the system's homogeneous differential equation: $\dot{\mathbf{\Phi}}(t,t_0) = \mathbf{A}\,\mathbf{\Phi}(t,t_0)$.

A.6 Application of the Laplace or the *z*-Transform

The transfer matrix of MIMO systems is obtained, as in the SISO case, by applying to the state space representation the Laplace or the z-transform, as illustrated in Table A.5.

Table A.5 Application of the Laplace or the z transform.

Continuous case	Discrete case
Resolvent matrix:	
$\mathcal{L}\left[e^{\mathbf{A}t}\right]=\mathcal{L}[\mathbf{\Phi}(t)]=(s\mathbf{I}-\mathbf{A})^{-1}$	$\mathcal{Z}[\mathbf{\Phi}(kT)]=\mathcal{Z}\left[\mathbf{\Phi}^{k}\right]=(z\mathbf{I}-\mathbf{\Phi})^{-1}z$
Transfer matrix:	
$\mathbf{Y}(s)=\mathbf{G}(s)\mathbf{U}(s)$	$\mathbf{Y}(z)=\mathbf{G}(z)\mathbf{U}(z)$
$\mathbf{G}(s)=\mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}+\mathbf{D}$	$\mathbf{G}(z)=\mathbf{C}(z\mathbf{I}-\mathbf{\Phi})^{-1}\mathbf{\Gamma}+\mathbf{D}$
Case of SISO systems: transfer function:	
$G(s)=\dfrac{\det\begin{pmatrix} s\mathbf{I}-\mathbf{A} & -\mathbf{b} \\ \mathbf{c}^{\mathrm{T}} & d \end{pmatrix}}{\det(s\mathbf{I}-\mathbf{A})}$	$G(z)=\dfrac{\det\begin{pmatrix} z\mathbf{I}-\mathbf{\Phi} & -\boldsymbol{\gamma} \\ \mathbf{c}^{\mathrm{T}} & d \end{pmatrix}}{\det(z\mathbf{I}-\mathbf{\Phi})}$

Every pole of $G(s)$, respectively of $G(z)$, is necessarily an *eigenvalue* of **A**, respectively of $\mathbf{\Phi}$, (see Sect. A.7.4), but the inverse is not true (example: case of an eigenvalue which cancels also the numerator).

Case of MIMO systems: poles and zeros

The eigenvalues (or poles, by language abuse) of the plant (A.7), respectively (A.8), are as previously the roots of $\det(s\mathbf{I}-\mathbf{A})$, respectively of $\det(z\mathbf{I}-\mathbf{\Phi})$. The zeros of this plant are the values of s, respectively of z, for which the equation

$$\begin{pmatrix} \mathbf{A}-s\mathbf{I} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}\begin{pmatrix} \mathbf{x} \\ \mathbf{u} \end{pmatrix}=\mathbf{0} \qquad (A.27) \qquad\qquad \begin{pmatrix} \mathbf{\Phi}-z\mathbf{I} & \mathbf{\Gamma} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}\begin{pmatrix} \mathbf{x} \\ \mathbf{u} \end{pmatrix}=\mathbf{0} \qquad (A.28)$$

has a non trivial solution, thus, in the case where $p=q$, the values of s, respectively of z, solutions of

$$\begin{vmatrix} \mathbf{A}-s\mathbf{I} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{vmatrix}=0 \qquad (A.29) \qquad\qquad \begin{vmatrix} \mathbf{\Phi}-z\mathbf{I} & \mathbf{\Gamma} \\ \mathbf{C} & \mathbf{D} \end{vmatrix}=0$$

The first member matrix of (A.27) is also called *Rosenbrock system matrix* [Rop90]. If $s=\mu$, respectively $z=\mu$, is such a zero, the solution $\mathbf{x}_Z$, respectively $\mathbf{u}_Z$, of (A.27), respectively of (A.28), given by

$$\begin{pmatrix} \mathbf{A}-\mu\mathbf{I} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}\begin{pmatrix} \mathbf{x}_Z \\ \mathbf{u}_Z \end{pmatrix}=\mathbf{0}, \qquad \begin{pmatrix} \mathbf{\Phi}-\mu\mathbf{I} & \mathbf{\Gamma} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}\begin{pmatrix} \mathbf{x}_Z \\ \mathbf{u}_Z \end{pmatrix}=\mathbf{0},$$

is called *state vector direction*, respectively *control vector direction*, *associated with this zero*.

Remark A.1. Dynamic characterization of a zero. If $\mathbf{x}_0=\mathbf{x}_Z$ and if, for $t\geq 0$, $\mathbf{u}(t)=\mathbf{u}_Z\,e^{\mu\,t}$, then $\mathbf{x}(t)=\mathbf{x}_Z\,e^{\mu\,t}$, which makes that

$$\mathbf{y}(t)=\mathbf{Cx}+\mathbf{Du}=(\mathbf{Cx}_Z+\mathbf{Du}_Z)e^{\mu t}\equiv\mathbf{0},\quad \forall t\,.$$

A.7 Transformation to one of the Canonical Forms

Due to the formal similitude of the two cases, the rules and transformation theorems which follow will be given only for the continuous case

A.7.1 Basis Change in the State Representations

Assume that new coordinates (state variables) have been introduced by means of a transformation defined by a regular matrix **T** such that

$$\mathbf{x}=\mathbf{T}\,\mathbf{z}\,,$$

where **x** and **z** denote respectively the old and the new state vector. In the new basis, the state equations become

$$\dot{\mathbf{z}}=\hat{\mathbf{A}}\,\mathbf{z}+\hat{\mathbf{B}}\,\mathbf{u}$$
$$\mathbf{y}=\hat{\mathbf{C}}\,\mathbf{z}+\hat{\mathbf{D}}\,\mathbf{u}$$

with:

$$\hat{\mathbf{A}}=\mathbf{T}^{-1}\mathbf{AT},\quad \hat{\mathbf{B}}=\mathbf{T}^{-1}\mathbf{B},\quad \hat{\mathbf{C}}=\mathbf{CT},\quad \hat{\mathbf{D}}=\mathbf{D}\,. \tag{A.30}$$

Such a transformation is called a *similitude* transformation, the matrices **A** and $\hat{\mathbf{A}}$ are called *similar* matrices.

Characteristic equation invariance in this transformation. The characteristic equation $\det(\lambda\mathbf{I}-\mathbf{A})=0$ is invariant in a change of basis defined by a regular transformation matrix **T**. A similitude transformation does not modify thus the eigenvalues of a matrix.

A.7.2 Transformation to Controllability Canonical Form

The following rule and theorem apply to single input systems.

Rule A.1. In the controllability canonical form, the last row of the system matrix is composed of the coefficients of the characteristic equation (except the one of the highest power term), with opposite signs.

Theorem A.1. *If the system* $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}u$ *(*$\mathbf{A}$ *and* $\mathbf{b}$ *constant) is controllable, it is possible, by means of the transformation* $\mathbf{z} = \mathbf{V}\mathbf{x} = \mathbf{T}^{-1}\mathbf{x}$ *to put it in the controllability canonical form*

$$\dot{\mathbf{z}} = \mathbf{A}_C\,\mathbf{z} + \mathbf{b}_C\,u\,,$$

where:

$$\mathbf{A}_C = \mathbf{V}\mathbf{A}\mathbf{V}^{-1} = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 \\ -a_0 & -a_1 & \cdots & -a_{n-1} \end{pmatrix}, \quad \mathbf{b}_C = \mathbf{V}\mathbf{b} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}, \tag{A.31}$$

and where the inverse transformation matrix is given by

$$\mathbf{V} = \begin{pmatrix} \mathbf{q}_C^{\mathrm{T}} \\ \mathbf{q}_C^{\mathrm{T}}\mathbf{A} \\ \vdots \\ \mathbf{q}_C^{\mathrm{T}}\mathbf{A}^{n-1} \end{pmatrix}, \tag{A.32}$$

where $\mathbf{q}_C^{\mathrm{T}}$ *is the last row of the inverse controllability matrix,* $\mathbf{Q}_C^{-1}$*, and is therefore deduced, according to (A.12), from the following system of equations:*

$$\begin{cases} \mathbf{q}_C^{\mathrm{T}}\mathbf{b} & = 0 \\ \mathbf{q}_C^{\mathrm{T}}\mathbf{A}\mathbf{b} & = 0 \\ \vdots & \\ \mathbf{q}_C^{\mathrm{T}}\mathbf{A}^{n-2}\mathbf{b} & = 0 \\ \mathbf{q}_C^{\mathrm{T}}\mathbf{A}^{n-1}\mathbf{b} & = 1 \end{cases} \tag{A.33}$$

i.e. also

$$\mathbf{q}_C^{\mathrm{T}}\mathbf{Q}_C = \begin{pmatrix} 0 & \cdots & 0 & 1 \end{pmatrix}.$$

A.7.3 Transformation to Observability Canonical Form

The following rule and theorem apply to single output systems.

Rule A.2. In the observability canonical form, the last column of the system matrix is composed of the coefficients of the characteristic equation (except the one of the highest power term), with opposite signs.

Theorem A.2. *If the system* $\begin{cases} \dot{\mathbf{x}} = \mathbf{A}\,\mathbf{x} + \mathbf{b}\,u \\ y = \mathbf{c}^{\mathrm{T}}\,\mathbf{x} \end{cases}$ *(*$\mathbf{A}$*,* $\mathbf{b}$*,* $\mathbf{c}^{\mathrm{T}}$ *constant) is observable, it is possible, by means of the transformation* $\mathbf{x} = \mathbf{T}\,\mathbf{z}$*, to put it in the observability canonical form*

$$\begin{cases} \dot{\mathbf{z}} = \mathbf{A}_{\mathcal{O}}\,\mathbf{z} + \mathbf{b}_{\mathcal{O}}\,u \\ y = \mathbf{c}_{\mathcal{O}}^{\mathrm{T}}\,\mathbf{z} \end{cases}$$

where

$$\mathbf{A}_{\mathcal{O}} = \mathbf{T}^{-1}\,\mathbf{A}\,\mathbf{T} = \begin{pmatrix} 0 & \cdots & \cdots & -a_0 \\ 1 & \ddots & & -a_1 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & -a_{n-1} \end{pmatrix}, \quad \mathbf{c}_{\mathcal{O}}^{\mathrm{T}} = \mathbf{c}^{\mathrm{T}}\,\mathbf{T} = \begin{pmatrix} 0 & \cdots & 0 & 1 \end{pmatrix}, \tag{A.34}$$

and where the transformation matrix is given by

$$\mathbf{T} = \begin{pmatrix} \mathbf{q}_{\mathcal{O}} & \mathbf{A}\,\mathbf{q}_{\mathcal{O}} & \cdots & \mathbf{A}^{n-1}\,\mathbf{q}_{\mathcal{O}} \end{pmatrix}, \tag{A.35}$$

where $\mathbf{q}_{\mathcal{O}}$ *is the last column of the inverse observability matrix,* $\mathbf{Q}_{\mathcal{O}}^{-1}$*, and is therefore deduced, according to (A.16), from the following system of equations:*

$$\begin{cases} \mathbf{c}^{\mathrm{T}}\mathbf{q}_{\mathcal{O}} = 0 \\ \mathbf{c}^{\mathrm{T}}\mathbf{A}\,\mathbf{q}_{\mathcal{O}} = 0 \\ \quad\vdots \\ \mathbf{c}^{\mathrm{T}}\mathbf{A}^{n-2}\,\mathbf{q}_{\mathcal{O}} = 0 \\ \mathbf{c}^{\mathrm{T}}\mathbf{A}^{n-1}\,\mathbf{q}_{\mathcal{O}} = 1 \end{cases} \tag{A.36}$$

i.e. also:

$$\mathbf{Q}_{\mathcal{O}}\,\mathbf{q}_{\mathcal{O}} = (0 \quad \cdots \quad 0 \quad 1)^{\mathrm{T}}.$$

A.7.4 Transformation to Diagonal Form

If the eigenvalues of the $(n \times n)$ matrix $\mathbf{A}$ are all distinct, the system $\begin{cases} \dot{\mathbf{x}} = \mathbf{A}\,\mathbf{x} + \mathbf{B}\,\mathbf{u} \\ \mathbf{y} = \mathbf{C}\,\mathbf{x} + \mathbf{D}\,\mathbf{u} \end{cases}$ can be represented in the new basis by the equations

$$\begin{cases} \dot{\mathbf{x}}^* = \mathbf{\Lambda}\,\mathbf{x}^* + \hat{\mathbf{B}}\,\mathbf{u} \\ \mathbf{y} = \hat{\mathbf{C}}\,\mathbf{x}^* + \hat{\mathbf{D}}\,\mathbf{u} \end{cases} \tag{A.37}$$

where, due to their particular importance, we have denoted by $\mathbf{x}^*$ instead of $\mathbf{z}$ the new state vector coordinates, also called *modal coordinates*, and where

$$\hat{\mathbf{A}} = \mathbf{T}^{-1}\mathbf{A}\,\mathbf{T} = \mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & \cdots & \cdots & 0 \\ \vdots & \lambda_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & \lambda_n \end{pmatrix}, \tag{A.38}$$

the matrices $\hat{\mathbf{B}}$, $\hat{\mathbf{C}}$ and $\hat{\mathbf{D}}$ being given by (A.30).

The *eigenvalues* $\lambda_1,\ldots,\lambda_n$ of $\mathbf{A}$ and its *eigenvectors* $\mathbf{v}_1,\ldots,\mathbf{v}_n$ are defined by the relation $\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i$, thus also

$$(\lambda_i\,\mathbf{I} - \mathbf{A})\mathbf{v}_i = 0\,. \tag{A.39}$$

Solving the *characteristic equation* of $\mathbf{A}$, i.e.

$$\det(\lambda_i\,\mathbf{I} - \mathbf{A}) = 0\,, \tag{A.40}$$

yields the eigenvalues λ_i, to which correspond thus the non trivial solutions $\mathbf{v}_i$ of (A.39), i.e. the eigenvectors of $\mathbf{A}$.

The new basis is composed of these eigenvectors, and the transformation matrix $\mathbf{T}$, called here *modal matrix*, is given by

$$\mathbf{T} = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nn} \end{pmatrix} = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{pmatrix}. \tag{A.41}$$

The *left eigenvectors* of **A**, by opposition to the above eigenvectors or *right* eigenvectors, are the vectors $\mathbf{w}_i$ which satisfy

$$\mathbf{w}_i^{\mathrm{T}}\mathbf{A} = \lambda_i \mathbf{w}_i^{\mathrm{T}}, \tag{A.42}$$

thus also

$$\mathbf{w}_i^{\mathrm{T}}(\lambda_i \mathbf{I} - \mathbf{A}) = 0. \tag{A.43}$$

It should be noted that the rows of the inverse transformation matrix $\mathbf{T}^{-1}$ are equal, apart from a multiplying factor, to the transposes of the left eigenvectors, $\mathbf{w}_i^{\mathrm{T}}$. Indeed, (A.38) can also be written $\mathbf{T}^{-1}\mathbf{A} = \mathbf{\Lambda}\mathbf{T}^{-1}$, therefore, denoting by $\mathbf{t}_i^{\mathrm{T}}$ the *i*-th row of $\mathbf{T}^{-1}$,

$$\begin{pmatrix} \mathbf{t}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{t}_n^{\mathrm{T}} \end{pmatrix} \mathbf{A} = \mathbf{\Lambda} \begin{pmatrix} \mathbf{t}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{t}_n^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix} \begin{pmatrix} \mathbf{t}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{t}_n^{\mathrm{T}} \end{pmatrix}.$$

The *i*-th row of this matrix equation, $\mathbf{t}_i^{\mathrm{T}}\mathbf{A} = \lambda_i \mathbf{t}_i^{\mathrm{T}}$, shows with (A.42) that

$$\mathbf{t}_i^{\mathrm{T}} = \mathbf{w}_i^{\mathrm{T}}. \tag{A.44}$$

Remark A.2. The left eigenvectors of a matrix **A** are right eigenvectors of its transpose, since (A.43) implies that $(\lambda_i \mathbf{I} - \mathbf{A}^{\mathrm{T}})\mathbf{w}_i = 0$.

A.7.5 Transient Response of a System as a Function of its Eigenmodes

The transient response of a linear time-invariant system to an initial condition $\mathbf{x}(t_0) = \mathbf{x}_0$ and an input excitation $\mathbf{u}(t)$, given by (A.18), involves the matrix exponential $e^{\mathbf{A}t}$, which can be expressed in a particularly simple way by using the diagonal representation. Indeed, thanks to (A.21) and (A.38) we can also write

$$e^{\mathbf{A}t} = \mathbf{T}e^{\mathbf{T}^{-1}\mathbf{A}t\mathbf{T}}\mathbf{T}^{-1} = \mathbf{T}e^{\mathbf{\Lambda}t}\mathbf{T}^{-1}.$$

Since $\mathbf{\Lambda} t$ is also a diagonal matrix, $e^{\mathbf{\Lambda} t}$ is it too and is composed of the exponentials of the diagonal elements of $\mathbf{\Lambda} t$.

Involving the right and left eigenvectors of **A**, thus the columns of **T** and the rows of $\mathbf{T}^{-1}$, we can write

$$e^{\mathbf{A}t} = \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{pmatrix} \begin{pmatrix} e^{\lambda_1 t} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{\lambda_n t} \end{pmatrix} \begin{pmatrix} \mathbf{w}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{w}_n^{\mathrm{T}} \end{pmatrix} = \sum_{i=1}^{n} \mathbf{v}_i e^{\lambda_i t} \mathbf{w}_i^{\mathrm{T}} \,,$$

which yields, from (A.18), the following system response

$$\mathbf{x}(t) = \sum_{i=1}^{n} \underbrace{\mathbf{v}_i \, e^{\lambda_i t}}_{eigenmode} \, \underbrace{\mathbf{w}_i^{\mathrm{T}} \mathbf{x}_0}_{\substack{scalar \\ product}} + \sum_{i=1}^{n} \mathbf{v}_i \int_0^t e^{\lambda_i (t-\tau)} \mathbf{w}_i^{\mathrm{T}} \, \mathbf{B} \, \mathbf{u}(\tau) \, d\tau \,. \tag{A.45}$$

This expression reveals the system *eigenmodes*, composed of an *eigendirection* $\mathbf{v}_i$ and a factor, which decays exponentially with a time constant equal to the inverse absolute value of the eigenvalue. These modes are involved in both the free response and the forced response, where right and left eigenvectors play complementary roles [Duc01]. More precisely,

- the eigenvector $\mathbf{v}_i$ determines the distribution of the corresponding mode $e^{\lambda_i t}$ on the components of the state vector: if e.g. the *j*-th component of $\mathbf{v}_i$ vanishes, the mode $e^{\lambda_i t}$ will not appear on the *j*-th component of $\mathbf{x}(t)$;
- the transpose left eigenvector $\mathbf{w}_i^{\mathrm{T}}$ determines the way the mode $e^{\lambda_i t}$ is excited by the different components of the initial state $\mathbf{x}_0$ and the row vector $\mathbf{w}_i^{\mathrm{T}} \mathbf{B}$ the way this mode is excited by the different components of the input $\mathbf{u}(t)$. Denoting e.g. by c_i the scalar product $\mathbf{w}_i^{\mathrm{T}} \mathbf{x}_0$, the initial state $\mathbf{x}_0$, as deduced from (A.45) written for $t = t_0 = 0$, can be written $\mathbf{x}_0 = \sum_{i=1}^{n} c_i \mathbf{v}_i$.

The c_i's are thus nothing else than the components of $\mathbf{x}_0$ in the eigenvector basis. In particular,

1. if $\mathbf{x}_0$ is collinear to $\mathbf{v}_j$, only the mode $e^{\lambda_j t}$ will appear in the free response;
2. if $\mathbf{x}_0$ is orthogonal to $\mathbf{w}_i$, the mode $e^{\lambda_i t}$ will not appear in the free response.

B Complements of Matrix Calculus

B.1 Trace of a Matrix

Definition B.1. If **A** is a square matrix of size $(n \times n)$, the trace of **A**, denoted $\mathrm{tr}(\mathbf{A})$, is the sum of its diagonal elements:

$$\mathbf{A} = (a_{ij}) \quad \Rightarrow \quad \mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii} \; .$$

Property B.1. If **A** is an $(n \times m)$ matrix and **B** an $(m \times n)$ matrix,

$$\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA}) \, . \tag{B.1}$$

Proof: $\mathrm{tr}(\mathbf{AB}) = \sum_{i=1}^{n} (\mathbf{AB})_{ii} = \sum_{i=1}^{n} \sum_{k=1}^{m} a_{ik} b_{ki} = \sum_{k=1}^{m} \sum_{i=1}^{n} b_{ki} a_{ik} = \sum_{k=1}^{m} (\mathbf{BA})_{kk} = \mathrm{tr}(\mathbf{BA})$.

Special case: *m* = 1. If **a** and **b** are two vectors of dimension *n*, the trace of a scalar being this scalar itself,

$$\mathrm{tr}(\mathbf{ab}^{\mathrm{T}}) = \mathbf{b}^{\mathrm{T}}\mathbf{a} \, . \tag{B.2}$$

Application. By letting $\mathbf{a} = \mathbf{Mx}$, where **M** is an $(n \times n)$ matrix, and $\mathbf{b} = \mathbf{x}$ of dimension *n*,

$$\mathbf{x}^{\mathrm{T}}\mathbf{Mx} = \mathrm{tr}(\mathbf{Mx} \cdot \mathbf{x}^{\mathrm{T}}) \, .$$

B.2 Matrix Inverses

After recalling the definition and calculation of the inverse of a matrix, a few useful inversion formulas will be given.

E. Ostertag, *Mono- and Multivariable Control and Estimation,*
DOI 10.1007/978-3-642-13734-1_8, © Springer-Verlag Berlin Heidelberg 2011

B.2.1 Inversion of a Matrix

If $\mathbf{A}$ is a square matrix, with $\det(\mathbf{A}) \neq 0$, it admits an inverse defined by $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ and calculated by

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})}\operatorname{adj}(\mathbf{A}), \quad \text{with } \left(\operatorname{adj}(\mathbf{A})\right)_{i,j} = (-1)^{i+j}\det(\mathbf{A}_{ji}),$$

where $\mathbf{A}_{ij}$ is the matrix obtained by deleting the i^{th} row and the j^{th} column from **A**. If $\mathbf{A}^{-1}$ exists, **A** is said *invertible* or *regular*, and $\det(\mathbf{A}^{-1}) = 1/\det(\mathbf{A})$.

Unitary matrix. A real matrix **U** is called *unitary*, if $\mathbf{U}^{\mathrm{T}} = \mathbf{U}^{-1}$. Thus $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$.

B.2.2 Matrix Inversion Lemma

If the matrices **A**, **C** and $\mathbf{A} + \mathbf{BCD}$ are invertible,

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}. \tag{B.3}$$

Proof. Let us left multiply the two members of this equation by $\mathbf{A} + \mathbf{BCD}$:

$$\begin{aligned}\mathbf{I} &= \mathbf{I} + \mathbf{BCDA}^{-1} - \mathbf{B}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1} - \mathbf{BCDA}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1} \\ &= \mathbf{I} + \mathbf{BCDA}^{-1} - \mathbf{B}(\mathbf{I} + \mathbf{CDA}^{-1}\mathbf{B})(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}\end{aligned}$$

By left factorizing **BC** in the third term of the second member, we get for it

$$\begin{aligned}\mathbf{I} + \mathbf{BCDA}^{-1} - \mathbf{BC}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1} \\ = \mathbf{I} + \mathbf{BCDA}^{-1} - \mathbf{BCDA}^{-1} = \mathbf{I}.\end{aligned}$$

B.2.3 Inverses of Some Particular Partitioned Matrices

First case: if the matrices **A** and **D** are invertible,

$$\begin{pmatrix}\mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{D}\end{pmatrix}^{-1} = \begin{pmatrix}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{BD}^{-1} \\ \mathbf{0} & \mathbf{D}^{-1}\end{pmatrix}. \tag{B.4}$$

Proof. By verification that the product of this matrix with its non-inverted original is the identity matrix, knowing that the inverse of a matrix is unique.

Second case: if the matrices **B** and **C** are invertible,

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{0} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{0} & \mathbf{C}^{-1} \\ \mathbf{B}^{-1} & -\mathbf{B}^{-1}\mathbf{A}\mathbf{C}^{-1} \end{pmatrix}. \tag{B.5}$$

Proof. Similar to the previous case.

B.3 Matrix Analysis

B.3.1 Differentiation of a Scalar with Respect to a Vector (Gradient)

Definition B.2. If y is a scalar function of n variables x_i, considered as the components of a column vector **x** (respectively of a row vector $\mathbf{x}^{\mathrm{T}}$), the expression

$$\frac{dy}{d\mathbf{x}} \quad \left(\text{respectively } \frac{dy}{d\mathbf{x}^{\mathrm{T}}}\right)$$

denotes a column vector (respectively a row vector) of components $\partial y / \partial x_i$:

$$\frac{dy}{d\mathbf{x}} = \begin{pmatrix} \dfrac{\partial y}{\partial x_1} \\ \vdots \\ \dfrac{\partial y}{\partial x_n} \end{pmatrix} = \boldsymbol{\nabla} y, \tag{B.6}$$

$$\frac{dy}{d\mathbf{x}^{\mathrm{T}}} = \begin{pmatrix} \dfrac{\partial y}{\partial x_1} & \cdots & \dfrac{\partial y}{\partial x_n} \end{pmatrix} = \left(\frac{dy}{d\mathbf{x}}\right)^{\mathrm{T}}. \tag{B.7}$$

Property B.2: If **a** and **x** are vectors of dimension n,

$$\frac{d}{d\mathbf{x}}(\mathbf{a}^{\mathrm{T}}\mathbf{x}) = \frac{d}{d\mathbf{x}}(\mathbf{x}^{\mathrm{T}}\mathbf{a}) = \mathbf{a}\,. \tag{B.8}$$

Proof. For the first part of this equation,

$$\frac{d}{d\mathbf{x}}(\mathbf{a}^{\mathrm{T}}\mathbf{x}) = \frac{d}{d\mathbf{x}}(a_1x_1 + \ldots + a_nx_n) = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = \mathbf{a}\,.$$

Furthermore, the transpose of a scalar being this scalar itself, $\mathbf{x}^{\mathrm{T}} \cdot \mathbf{a} = (\mathbf{x}^{\mathrm{T}} \cdot \mathbf{a})^{\mathrm{T}} = \mathbf{a}^{\mathrm{T}} \cdot \mathbf{x}$, which proves the second half of (B.8).

Property B.3: $$\frac{d}{d\mathbf{x}}(\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}) = (\mathbf{A} + \mathbf{A}^{\mathrm{T}})\,\mathbf{x}\,. \tag{B.9}$$

Proof. By comparison with the scalar situation, where

$$\frac{d}{dx}(uv) = \frac{du}{dx}v + u\frac{dv}{dx} = \frac{d}{dx}(uv)_{v=\text{constant}} + \frac{d}{dx}(uv)_{u=\text{constant}}\,,$$

the following applies here:

$$\frac{d}{d\mathbf{x}}(\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}) = \frac{d}{d\mathbf{x}}(\mathbf{x}^{\mathrm{T}}\mathbf{a})\bigg|_{\mathbf{a}=\mathbf{A}\mathbf{x}} + \frac{d}{d\mathbf{x}}(\mathbf{b}^{\mathrm{T}}\mathbf{x})\bigg|_{\mathbf{b}^{\mathrm{T}}=\mathbf{x}^{\mathrm{T}}\mathbf{A}} = \mathbf{A}\mathbf{x} + \mathbf{A}^{\mathrm{T}}\mathbf{x}\,.$$

Remark B.1. If **A** is symmetric,

$$\frac{d}{d\mathbf{x}}(\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}) = 2\mathbf{A}\mathbf{x}\,. \tag{B.10}$$

Special case $\mathbf{A} = \mathbf{I}$:

$$\frac{d}{d\mathbf{x}}(\mathbf{x}^{\mathrm{T}}\mathbf{x}) = 2\mathbf{x}\,. \tag{B.11}$$

B.3.2 Differentiation of a Vector with Respect to a Vector

Definition B.3. If **y** is an $(m \times 1)$ vector, the elements of which are functions of n variables x_i, considered as the components of a $(1 \times n)$ row vector, $\mathbf{x}^{\mathrm{T}}$, the expression $d\mathbf{y}/d\mathbf{x}^{\mathrm{T}}$ denotes an $(m \times n)$ matrix, with elements $\partial y_i/\partial x_j$:

$$\frac{d\mathbf{y}}{d\mathbf{x}^{\mathrm{T}}}=\begin{pmatrix}\dfrac{\partial y_1}{\partial x_1} & \cdots & \dfrac{\partial y_1}{\partial x_n}\\ \vdots & & \vdots\\ \dfrac{\partial y_m}{\partial x_1} & \cdots & \dfrac{\partial y_m}{\partial x_n}\end{pmatrix}. \tag{B.12}$$

Property B.4:

$$\frac{d\mathbf{y}^{\mathrm{T}}}{d\mathbf{x}}=\left(\frac{\partial y_j}{\partial x_i}\right)=\left(\frac{d\mathbf{y}}{d\mathbf{x}^{\mathrm{T}}}\right)^{\mathrm{T}}. \tag{B.13}$$

Remark B.2. If **x** and **y** have the same dimension, $\det(d\mathbf{y}/d\mathbf{x}^{\mathrm{T}})$ is called *Jacobian determinant*, or simply *Jacobian*, of $\mathbf{y}$.

Property B.5:

$$\frac{d}{d\mathbf{x}^{\mathrm{T}}}(\mathbf{A}\mathbf{x})=\frac{d}{d\mathbf{x}}(\mathbf{x}^{\mathrm{T}}\mathbf{A})=\mathbf{A}. \tag{B.14}$$

Proof. Let $\mathbf{y}=\mathbf{A}\mathbf{x}$, with $\mathbf{A}\in\mathbb{R}^{m\times n}$. According to the definition equation (B.12),

$$\frac{d\mathbf{y}}{d\mathbf{x}^{\mathrm{T}}}=\frac{d}{d\mathbf{x}^{\mathrm{T}}}\begin{pmatrix}a_{11}x_1+\;\ldots\;+a_{1n}x_n\\ \vdots\\ a_{m1}x_1+\;\ldots\;+a_{mn}x_n\end{pmatrix}=\begin{pmatrix}a_{11} & \cdots & a_{1n}\\ \vdots & & \vdots\\ a_{m1} & \cdots & a_{mn}\end{pmatrix}=\mathbf{A}.$$

For the second half of the relation, it holds that

$$\frac{d}{d\mathbf{x}}(\mathbf{x}^{\mathrm{T}}\mathbf{A})=\left[\frac{d}{d\mathbf{x}^{\mathrm{T}}}(\mathbf{x}^{\mathrm{T}}\mathbf{A})^{\mathrm{T}}\right]^{\mathrm{T}}=\left[\frac{d}{d\mathbf{x}^{\mathrm{T}}}(\mathbf{A}^{\mathrm{T}}\mathbf{x})\right]^{\mathrm{T}}=(\mathbf{A}^{\mathrm{T}})^{\mathrm{T}}.$$

Remark B.3. Special case $\mathbf{A}=\mathbf{I}$:

$$\frac{d\mathbf{x}}{d\mathbf{x}^{\mathrm{T}}}=\frac{d\mathbf{x}^{\mathrm{T}}}{d\mathbf{x}}=\mathbf{I}. \tag{B.15}$$

Property B.6: Case of a dot product containing a vector function of x. Consider the vectors **x**, of dimension *n*, and **y**, of dimension *m*, function of **x**, and the constant vector **a** of dimension *m*. The following differentiation formulae hold:

$$\frac{d}{d\mathbf{x}}(\mathbf{a}^{\mathrm{T}}\mathbf{y})=\frac{d}{d\mathbf{x}}(\mathbf{y}^{\mathrm{T}}\mathbf{a})=\frac{d\mathbf{y}^{\mathrm{T}}}{d\mathbf{x}}\mathbf{a}. \tag{B.16}$$

Proof. Since $\mathbf{a}^{\mathrm{T}}\mathbf{y}$ is a scalar, we have, according to (B.6),

$$\frac{d}{d\mathbf{x}}(\mathbf{a}^{\mathrm{T}}\mathbf{y}) = \frac{d}{d\mathbf{x}}(a_1 y_1 + \ \ldots \ + a_m y_m) = \begin{pmatrix} a_1 \dfrac{\partial y_1}{\partial x_1} + \ \ldots \ + a_m \dfrac{\partial y_m}{\partial x_1} \\ \vdots \\ a_1 \dfrac{\partial y_1}{\partial x_n} + \ \ldots \ + a_m \dfrac{\partial y_m}{\partial x_n} \end{pmatrix}.$$

The obtained expression can also be written as

$$\frac{d}{d\mathbf{x}}(\mathbf{a}^{\mathrm{T}}\mathbf{y}) = \begin{pmatrix} \dfrac{\partial y_1}{\partial x_1} & \cdots & \dfrac{\partial y_m}{\partial x_1} \\ \vdots & & \vdots \\ \dfrac{\partial y_1}{\partial x_n} & \cdots & \dfrac{\partial y_m}{\partial x_n} \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} = \frac{d\mathbf{y}^{\mathrm{T}}}{d\mathbf{x}} \mathbf{a} \,,$$

according to (B.12) defining the derivative of a vector with respect to a vector.

Furthermore, $\mathbf{y}^{\mathrm{T}}\mathbf{a} = \mathbf{a}^{\mathrm{T}}\mathbf{y}$, which proves the second half of (B.16).

B.3.3 Differentiation of a Scalar with Respect to a Matrix

Definition B.4. If y is a scalar function of $m \times n$ variables x_{ij} , considered as the elements of a matrix **X**, the expression $dy/d\mathbf{X}$ denotes a matrix, with elements $\partial y/\partial x_{ij}$:

$$\frac{dy}{d\mathbf{X}} = \begin{pmatrix} \dfrac{\partial y}{\partial x_{11}} & \cdots & \dfrac{\partial y}{\partial x_{1n}} \\ \vdots & & \vdots \\ \dfrac{\partial y}{\partial x_{m1}} & \cdots & \dfrac{\partial y}{\partial x_{mn}} \end{pmatrix}. \tag{B.17}$$

The following properties of this subsection will hold only in the case **X** is a square matrix, of size $(n \times n)$.

Property B.7:

$$\frac{d}{d\mathbf{X}} \operatorname{tr}(\mathbf{X}) = \mathbf{I} \,. \tag{B.18}$$

Proof:

$$\frac{d}{d\mathbf{X}}\operatorname{tr}(\mathbf{X}) = \frac{d}{d\mathbf{X}}\left(\sum_i x_{ii}\right) = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & 1 \end{pmatrix} = \mathbf{I}\,.$$

Property B.8:

$$\frac{d}{d\mathbf{X}}\operatorname{tr}(\mathbf{A}\mathbf{X}^{\mathrm{T}}) = \mathbf{A}\,. \tag{B.19}$$

Proof:

$$\frac{d}{d\mathbf{X}}\operatorname{tr}(\mathbf{A}\mathbf{X}^{\mathrm{T}}) = \frac{d}{d\mathbf{X}}\left(\sum_i\sum_k a_{ik}x_{ik}\right) = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} = \mathbf{A}\,.$$

Property B.9:

$$\frac{d}{d\mathbf{X}}\operatorname{tr}(\mathbf{X}\mathbf{A}\mathbf{X}^{\mathrm{T}}) = \mathbf{X}(\mathbf{A}+\mathbf{A}^{\mathrm{T}})\,. \tag{B.20}$$

Proof. Following a similar approach to the proof of (B.9), we can write

$$\frac{d}{d\mathbf{X}}\operatorname{tr}(\mathbf{X}\mathbf{A}\mathbf{X}^{\mathrm{T}}) = \frac{d}{d\mathbf{X}}\operatorname{tr}\left[\mathbf{X}\cdot \underbrace{(\mathbf{A}\mathbf{X}^{\mathrm{T}})}_{\text{constant}}\right] + \frac{d}{d\mathbf{X}}\operatorname{tr}\left[\underbrace{(\mathbf{X}\mathbf{A})}_{\text{constant}}\cdot\mathbf{X}^{\mathrm{T}}\right].$$

Using (B.19) and its transpose, the previous expression becomes

$$\frac{d}{d\mathbf{X}}\operatorname{tr}(\mathbf{X}\mathbf{A}\mathbf{X}^{\mathrm{T}}) = (\mathbf{A}\mathbf{X}^{\mathrm{T}})^{\mathrm{T}} + \mathbf{X}\mathbf{A} = \mathbf{X}\mathbf{A}^{\mathrm{T}} + \mathbf{X}\mathbf{A}\,.$$

Remark B.4. If **A** is symmetric,

$$\frac{d}{d\mathbf{X}}\operatorname{tr}(\mathbf{X}\mathbf{A}\mathbf{X}^{\mathrm{T}}) = 2\mathbf{X}\mathbf{A}\,. \tag{B.21}$$

B.4 Linear Algebra

Here only main results will be presented. For more details the reader is referred to textbooks such as [Mey01] for basic Linear Algebra, or [ZhDG96] for more specific properties used in the Control Systems context.

Let some matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ be a linear transformation from the vector space $\mathbb{R}^n$ to the vector space $\mathbb{R}^m$.

Definition B.5. The *column space* $C(\mathbf{A})$ of **A**, sometimes also called the *image* or *range* of **A**, is the set of all linear combinations of the columns of **A**. It is therefore a subspace of $\mathbb{R}^m$, said to be *spanned* by the columns of **A**. It can be considered also as the following set:

$$C(\mathbf{A}) := \left\{ \mathbf{y} \in \mathbb{R}^m : \mathbf{y} = \mathbf{A}\mathbf{x}, \quad \mathbf{x} \in \mathbb{R}^n \right\}.$$

Definition B.6. The *null space* $N(\mathbf{A})$ of **A**, sometimes also called the *kernel* $\mathrm{Ker}(\mathbf{A})$ of **A**, is the set of all solutions of $\mathbf{A}\mathbf{x} = \mathbf{0}$, i.e.

$$N(\mathbf{A}) := \left\{ \mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} = \mathbf{0} \right\}.$$

It is therefore a subspace of $\mathbb{R}^n$.

Definition B.7. The *orthogonal complement* of a subspace $S \subset \mathbb{R}^n$ is defined by

$$S^{\perp} := \left\{ \mathbf{y} \in \mathbb{R}^n : \mathbf{y}^{\mathrm{T}}\mathbf{x} = 0 \text{ for all } \mathbf{x} \in S \right\}.$$

Property B.10. It is easy to show that $\dim\left[C(\mathbf{A})\right] + \dim\left[N(\mathbf{A})\right] = n$, where the dimension of a subspace *S*, noted $\dim(S)$, is the size of a basis of this subspace. Moreover, the following holds:

$$N(\mathbf{A}) = \left[C(\mathbf{A}^{\mathrm{T}})\right]^{\perp} \quad \Leftrightarrow \quad N(\mathbf{A}) \perp C(\mathbf{A}^{\mathrm{T}}),$$

in other words, the null space of **A** is the set of vectors orthogonal to all rows of **A**, which are the columns of $\mathbf{A}^{\mathrm{T}}$.

Definition B.8. The system of *m* linear equations $\mathbf{A}\mathbf{x} = \mathbf{y}$ in the *n* unknowns $x_1, \ldots, x_n$, the components of **x**, has a solution if and only if $\mathbf{y} \in C(\mathbf{A})$.
If **A** is square, of size $(n \times n)$, this solution is unique if and only if $\mathrm{rank}(\mathbf{A}) = n$.

Definition B.9. The *rank* of **A**, denoted $\mathrm{rank}(\mathbf{A})$, is the size of the largest non vanishing determinant contained in **A**. It is therefore equal to the maximal number of independent rows or columns, i.e.,

$$\mathrm{rank}(\mathbf{A}) = \dim\left[C(\mathbf{A})\right].$$

B.5 Singular Value Decomposition

An interesting tool in matrix analysis, which is very useful in systems analysis and control in the MIMO case, is the *Singular Value Decomposition (SVD).*

Theorem B.1. *Let* $\mathbf{A} \in \mathbb{R}^{m \times n}$ *. There exist unitary matrices*

$$\mathbf{U} = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_m \end{pmatrix} \in \mathbb{R}^{m \times m}, \quad \mathbf{U}^\mathrm{T}\mathbf{U} = \mathbf{I}_m$$
$$\mathbf{V} = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{pmatrix} \in \mathbb{R}^{n \times n}, \quad \mathbf{V}^\mathrm{T}\mathbf{V} = \mathbf{I}_n$$

such that

$$\mathbf{A} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^\mathrm{T}, \quad \mathbf{S} = \begin{pmatrix} \mathbf{S}_1 & \mathbf{0}_{p \times (n-p)} \\ \mathbf{0}_{(m-p) \times p} & \mathbf{0}_{(m-p) \times (n-p)} \end{pmatrix} \tag{B.22}$$

where

$$\mathbf{S}_1 = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p \end{pmatrix}$$

and

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0, \quad p = \min\{m, n\}.$$

The σ_i are the *singular values* of **A**. The vectors $\mathbf{u}_i$ and $\mathbf{v}_i$, called respectively *left singular vectors* and *right singular vectors* of **A**, satisfy respectively

$$\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i$$
$$\mathbf{A}^\mathrm{T}\mathbf{u}_i = \sigma_i \mathbf{v}_i.$$

Connection with eigenvalues and eigenvectors. The above equations can be written as

$$\mathbf{A}^\mathrm{T}\mathbf{A}\,\mathbf{v}_i = \sigma_i^2 \mathbf{v}_i$$
$$\mathbf{A}\,\mathbf{A}^\mathrm{T}\mathbf{u}_i = \sigma_i^2 \mathbf{u}_i.$$

Thus, σ_i^2 is an eigenvalue of $\mathbf{A}\,\mathbf{A}^\mathrm{T}$ and $\mathbf{A}^\mathrm{T}\mathbf{A}$, and $\mathbf{u}_i$, respectively $\mathbf{v}_i$, are eigenvectors of $\mathbf{A}\,\mathbf{A}^\mathrm{T}$, respectively $\mathbf{A}^\mathrm{T}\mathbf{A}$.

Maximum and minimum singular values. It is common practice to denote

$$\bar{\sigma}(\mathbf{A}) = \sigma_1 = \text{the largest singular value of } \mathbf{A}$$
$$\underline{\sigma}(\mathbf{A}) = \sigma_p = \text{ the smallest singular value of } \mathbf{A}$$

It can be shown that, for any eigenvector $\mathbf{v}_i$ of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$, we have $\|\mathbf{A}\,\mathbf{v}_i\| = \sigma_i \|\mathbf{v}_i\|$, where $\|\cdot\|$ denotes the Euclidian norm of a vector, i.e. its geometric length $\sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$. Hence the following interesting interpretation of singular vectors:

- $\mathbf{v}_1$ $(\mathbf{v}_n)$ is the input, i.e. *controlling* direction with the highest (lowest) gain;
- $\mathbf{u}_1$ $(\mathbf{u}_m)$ is the output, i.e. *observing* direction with the highest (lowest) gain.

If **A** is square and invertible, (B.22) and the unitary property of **U** and **V** yield

$$\mathbf{A}^{-1} = \mathbf{V}\,\mathbf{S}^{-1}\mathbf{U}^{\mathrm{T}},$$

which implies that the singular values of $\mathbf{A}^{-1}$ are the inverses of those of **A**. In particular,

$$\bar{\sigma}(\mathbf{A}^{-1}) = \frac{1}{\underline{\sigma}(\mathbf{A})} \tag{B.23}$$

Condition number. The *condition number* of **A** is defined as

$$\gamma(\mathbf{A}) = \frac{\bar{\sigma}(\mathbf{A})}{\underline{\sigma}(\mathbf{A})} \geq 1\,.$$

According to this definition and to (B.23), $\gamma(\mathbf{A}^{-1}) = \gamma(\mathbf{A})$.

When the condition number is very large, **A** is said *ill conditioned.* A small change of one of its elements can change its rank (see below).

Connection with the rank of a matrix: lemma B.1. *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *and*

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = 0, \quad r \leq \min\{m,n\}$$

be its strictly positive singular values.

Then $\operatorname{rank}(\mathbf{A}) = r$, $N(\mathbf{A})$ *is spanned by* $\{\mathbf{v}_{r+1},\ldots,\mathbf{v}_n\}$ *and* $C(\mathbf{A}^{\mathrm{T}}) = [N(\mathbf{A})]^{\perp}$ *is spanned by* $\{\mathbf{v}_1,\ldots,\mathbf{v}_r\}$.

As mentioned by several authors, e.g. by [AlSa04], this is a safer way to test the rank of a matrix than Definition B.9. A very large condition number, indeed, may mean that the smallest singular value is so close to zero that the *practical* rank of **A** is $r-1$ instead of *r*. A striking example is given in [AlSa04], Appendix B.

Reduced singular value decomposition. Since **S** contains mainly zeros, a reduced SVD is sometimes introduced, in which most of these zeros are eliminated.

Theorem B.2. *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}$*, with* $\operatorname{rank}(\mathbf{A}) = r$ *. Then there exist unitary matrices* $\mathbf{U}_1 \in \mathbb{R}^{m\times r}$ *and* $\mathbf{V}_1 \in \mathbb{R}^{n\times r}$ *such that*

$$\mathbf{A} = \mathbf{U}_1 \boldsymbol{\Sigma} \mathbf{V}_1^{\mathrm{T}},$$

where

$$\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1,\ldots,\sigma_r),$$

and

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0 .$$

$\mathbf{U}_1$*, respectively* $\mathbf{V}_1$*, is obtained from* **U***, respectively* **V**, *by dropping its rightmost* $(m-r)$*, respectively* $(n-r)$*, columns.*

B.6 Positive Definite and Positive Semidefinite Matrices

Consider a matrix **A** of size $n\times n$ and symmetric.

Definition B.10. **A** is called *positive definite* if for any vector $\mathbf{x} \neq 0$ the quadratic form $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ is positive.

Definition B.11. **A** is called *positive semidefinite* (or *nonnegative definite*) if for any vector $\mathbf{x} \neq 0$ the quadratic form $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ is positive or equal to zero.

Criterion of positive definiteness (Sylvester criterion). A matrix **A** is positive definite, if and only if the determinants associated with all upper-left submatrices of **A** are positive, i.e.

$$a_{11} > 0 ; \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{vmatrix} > 0 ; \quad \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{vmatrix} > 0 ; \quad \cdots \; ; \quad \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{12} & a_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{1n} & \cdots & \cdots & a_{nn} \end{vmatrix} > 0 .$$

Criterion of positive semidefiniteness. A matrix **A** is positive semidefinite, if and only if the determinants associated with all submatrices of **A**, the diagonals of which belong to the diagonal of **A**, are positive or equal to zero, i.e., as illustrated here in the case of a 3×3 matrix:

$$a_{11}, a_{22}, a_{33} \geq 0\,; \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{vmatrix}, \begin{vmatrix} a_{11} & a_{13} \\ a_{13} & a_{33} \end{vmatrix}, \begin{vmatrix} a_{22} & a_{23} \\ a_{23} & a_{33} \end{vmatrix} \geq 0\,; \quad \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{vmatrix} \geq 0\,.$$

Equivalent criterion for symmetric matrices. A symmetric matrix **A** is positive definite if and only if all its eigenvalues are positive; it is positive semidefinite if and only if all its eigenvalues are positive or equal to zero (nonnegative).

Property B.11. Given an $m \times n$ matrix **D**, the symmetric $n \times n$ matrix $\mathbf{A} = \mathbf{D}^{\mathrm{T}}\mathbf{D}$ is positive semidefinite. It is positive definite if and only if **D** is of rank *n*.

Proof. Consider the vector **y** defined by $\mathbf{y} = \mathbf{D}\mathbf{x}$. Then the following holds: $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} = \mathbf{x}^{\mathrm{T}}\mathbf{D}^{\mathrm{T}}\mathbf{D}\mathbf{x} = \mathbf{y}^{\mathrm{T}}\mathbf{y} = \sum y_i^2 \geq 0$. The inequality of the last member of this expression becomes an equality if and only if $\mathbf{y} = 0$, thus $\mathbf{D}\mathbf{x} = 0$, which is impossible for $\mathbf{x} \neq 0$ if **D** is of rank *n*.

Square root of a matrix. Given a matrix **D** and a matrix **A** as defined in Property B.11, **D** is called a *square root* of **A**. If **D** is a square root having a number of rows equal to the rank of **A**, all other square roots are obtained by **DT**, where **T** is any matrix for which $\mathbf{T}\mathbf{T}^{\mathrm{T}} = \mathbf{I}$.

If **A** is positive semidefinite, there exists a matrix $\mathbf{A}^{1/2}$ which is a *symmetric square root* of **A**, and which is itself positive semidefinite. It satisfies $\mathbf{A} = \mathbf{A}^{1/2}\mathbf{A}^{1/2}$. Moreover, if **A** is regular, so is $\mathbf{A}^{1/2}$.

Other properties:

- If **A** and **B** are positive semidefinite, so is $\mathbf{A} + \mathbf{B}$.
- If one of them is positive definite, so is $\mathbf{A} + \mathbf{B}$.
- If **A** is an $n \times n$ positive semidefinite matrix and if **B** is an $m \times n$ matrix, then $\mathbf{B}\mathbf{A}\mathbf{B}^{\mathrm{T}}$ is positive semidefinite.

C Review of Probability Theory

C.1 Probability Theory and Random Processes in the Scalar Case

This appendix is a summary of the main concepts. For a detailed presentation, refer to text books, such as [Kay06], [ShBr88] or [BrSi75a and b].

C.1.1 Random Variable

C.1.1.1 Definitions

A random variable *X* is a quantity whose value *x* depends on the outcome of a random experiment. A random variable is said *continuous* if it takes its values *x* in $]-\infty,+\infty[$; it is said *discrete* if it can take a finite number, or a countably infinite number, of real distinct values: $x_0, x_1, \ldots$.

C.1.1.2 Probability Distributions and Main Characteristics

The basic concepts and properties of a single random variable are summarized in Table C.1 for both the continuous and the discrete situations.

The case of two or of *n* random variables (*random vectors*) will be discussed in subsequent sections.

E. Ostertag, *Mono- and Multivariable Control and Estimation,*
DOI 10.1007/978-3-642-13734-1_9, © Springer-Verlag Berlin Heidelberg 2011

Table C.1 Main concepts of random distributions for the continuous and discrete cases.

Continuous Case	Discrete Case
Distribution function: $F(x)=P\{X<x\}$	
Probability density function: $p(x)$:	Probability mass function: p_k :
$p(x)dx=P\{x\le X<x+dx\}$	$p_k=P\{X=x_k\}$
$p(x)\ge 0,\ \int_{-\infty}^{\infty} p(x)dx=1$	$p_k\ge 0,\ \sum_k p_k=1$
Expected value or (mathematical) expectation, or (statistical) mean, or 1st order moment:	
$\mu=\mathrm{E}\{X\}=\int_{-\infty}^{\infty} x\,p(x)\,dx$	$\mu=\mathrm{E}\{X\}=\sum_k p_k\,x_k$
Expectation of the random variable $Y=g(X)$, where $g(x)$ is a real function of the variable *x*:	
$\mathrm{E}\{g(X)\}=\int_{-\infty}^{\infty} g(x)\,p(x)\,dx$	$\mathrm{E}\{g(X)\}=\sum_k p_k\,g(x_k)$
Moment of order *n* = expectation of $g(X)=X^n$: $m_n=\mathrm{E}\{X^n\}$	
$m_n=\int_{-\infty}^{\infty} x^n\ p(x)\,dx$	$m_n=\sum_k p_k\,x_k^n$
Centered moment of order *n*: $\mathrm{E}\{(X-\mu)^n\}$	
Variance of *X*, or centered moment of second order:	
$\mathrm{Var}\{X\}=\mathrm{E}\{(X-\mu)^2\}$ (C.1)	
σ_X , or simply σ if there is no ambiguity, denotes the standard deviation of *X*, which is equal to the square root of the variance: $\sigma_X^2=\mathrm{Var}\{X\}$.	

C.1.1.3 Two Random Variables

The previous notions extend themselves easily to a set of two random variables, *X* and *Y*, defined with respect to the same experiment. Let us consider here only the case where they are continuous, and define $\mu_X=\mathrm{E}\{X\}$ and $\mu_Y=\mathrm{E}\{Y\}$.

Bivariate distribution function:

$$F_{XY}(x,y)=P\{X<x \text{ and } Y<y\}\ . \tag{C.2}$$

Joint probability density: function of two variables $p(x,y)$ such that

$$p(x,y)\,dx\,dy=P\{x\le X<x+dx,\,y\le Y<y+dy\}\,, \tag{C.3}$$

with $$p(x,y)\geq 0, \quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\,dx\,dy = 1\,.$$

Expectation of a function of *X* and *Y*, i.e. $g(X,Y)$:

$$\mathrm{E}\{g(X,Y)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\,p(x,y)\,dx\,dy\,.$$

Mixed moments: let $g(x,y)=x^r y^s$, where *r* and *s* are positive integers:

$$m_{r,s} = \mathrm{E}\left\{X^r Y^s\right\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^r y^s\,p(x,y)\,dx\,dy\,. \tag{C.4}$$

Central mixed moments:

$$\mathrm{E}\left\{(X-\mu_X)^r\,(Y-\mu_Y)^s\right\}. \tag{C.5}$$

Covariance: it is the centered moment of second order:

$$\mathrm{Cov}\{X,Y\} = \mathrm{E}\{(X-\mu_X)(Y-\mu_Y)\} = \mathrm{E}\{XY\} - \mu_X\mu_Y\,. \tag{C.6}$$

Correlation coefficient or simply *correlation:*

$$\rho(X,Y) = \frac{\mathrm{Cov}\{\mathrm{X},\mathrm{Y}\}}{\sqrt{\mathrm{Var}\{X\}\,\mathrm{Var}\{Y\}}} = \frac{\mathrm{Cov}\{\mathrm{X},\mathrm{Y}\}}{\sigma_X\ \sigma_Y}\,. \tag{C.7}$$

Definitions C.1. Two random variables X and Y are said

- *uncorrelated* if $\mathrm{Cov}\{\mathrm{X},\mathrm{Y}\}=0$; then, (see (C.6)), $\mathrm{E}\{XY\}=\mathrm{E}\{X\}\mathrm{E}\{Y\}$;
- *orthogonal* if $\mathrm{E}\{XY\}=0$;
- *(statistically) independent* if $p(x,y)=p(x)p(y)$.

If *X* and *Y* are independent, then, according to (C.4): $\mathrm{E}\{XY\}=\mathrm{E}\{X\}\mathrm{E}\{Y\}$. Consequently, according to (C.6): $\mathrm{Cov}\{X,Y\}=0$, i.e. the random variables *X* and *Y* are *uncorrelated.*

Conclusion: two *independent* random variables are thus *always uncorrelated.*

However, the inverse is *not true*. Two uncorrelated random variables are not necessarily independent, except if they have Gaussian distributions.

C.1.2 Random Function. Random, or Stochastic Process

Definition C.2. A function $X(t)$ of the independent variable *t* is a *random function* when, for any value of the independent variable, its value is random. A *random* (or *stochastic*) *process* characterizes the evolution, as a function of some parameter *t*, of a system whose behavior is the result of chance.

In the majority of situations, but not always, the considered parameter *t* is the time, either continuous, or discrete.

In practice, the random functions appear often as signals which are logged as functions of time. In the following, we will consider the set $\mathscr{E}_N$ of the values of a random function $X(t)$, defined by *N* records $x_i(t)$ (*N* realizations of the random experiment), where *N* can be finite, countably or uncountably infinite. Such a record $x_i(t)$ is also called a *sample*, or a *member function*.

C.1.3 Continuous-time Random Processes

C.1.3.1 Ensemble Characteristics of First Order

The value at some time *t* of the various samples is a random variable $X(t)$, from which are defined the following characteristics of the random process:

Distribution function: $F(x,t) = P\{X(t) < x\}$.

Probability density: function $p(x,t)$ such that $p(x,t)\,dx = P\{x \le X(t) < x + dx\}$

Process mean: $\mu_X(t) = \mathrm{E}\{X(t)\}$.

Process variance: $\mathrm{Var}\{X(t)\} = \mathrm{E}\left\{\left[X(t) - \mathrm{E}\{X(t)\}\right]^2\right\}$.

C.1.3.2 Ensemble Characteristics of Second Order

The observation of the various samples at two distinct times, t_1 and t_2, constitutes a two-dimensional random variable $\left[X(t_1), X(t_2)\right]$.

Distribution function: $F(x_1,x_2,t_1,t_2) = P\{X(t_1) < x_1 \text{ and } X(t_2) < x_2\}$.

Joint probability density: function $p(x_1,x_2,t_1,t_2)$ such that

$$p(x_1,x_2,t_1,t_2)\,dx_1 dx_2 = P\{x_1 \le X(t_1) < x_1 + dx_1 \text{ and } x_2 \le X(t_2) < x_2 + dx_2\}.$$

C.1.3.3 Stationarity

A stochastic process is said *stationary of second order*, or *wide sense stationary*, if its first and second order ensemble characteristics do not depend on the origin of time. In particular, its ensemble mean is constant: $\mu_X(t) = \mu_X$.

The main second order ensemble characteristics are collected in Table C.2, in the case of a single stochastic process $X(t)$, on the left side, and of two distinct stochastic processes, $X(t)$ and $Y(t)$, on the right side. In the case of stationary processes, these characteristics depend only on the difference between the two instants, $t_2 - t_1 = \tau$.

Table C.2 Second order ensemble characteristics of one and two continuous random processes.

One process: $X(t)$	**Two processes:** $X(t)$ **and** $Y(t)$
1. General case (arbitrary processes):	
Autocovariance function:	Cross-covariance function:
$r_{XX}(t_2,t_1) = \mathrm{Cov}\{X(t_2),X(t_1)\}$ $= \mathrm{E}\{[X(t_2)-\mu_X(t_2)][X(t_1)-\mu_X(t_1)]\}$	$r_{XY}(t_2,t_1) = \mathrm{Cov}\{X(t_2),Y(t_1)\}$ $= \mathrm{E}\{[X(t_2)-\mu_X(t_2)][Y(t_1)-\mu_Y(t_1)]\}$
Autocorrelation function:	Cross-correlation function:
$\varphi_{XX}(t_2,t_1) = \mathrm{E}\{X(t_2)X(t_1)\}$	$\varphi_{XY}(t_2,t_1) = \mathrm{E}\{X(t_2)Y(t_1)\}$
2. *Stationary* processes: $t_2 - t_1 = \tau$	
Autocovariance function:	Cross-covariance function:
$r_{XX}(\tau) = \mathrm{Cov}\{X(t+\tau),X(t)\}$	$r_{XY}(\tau) = \mathrm{Cov}\{X(t+\tau),Y(t)\}$
Autocorrelation function:	Cross-correlation function:
$\varphi_{XX}(\tau) = \mathrm{E}\{X(t+\tau)X(t)\}$	$\varphi_{XY}(\tau) = \mathrm{E}\{X(t+\tau)Y(t)\}$
3. *Centered stationary* processes: $\mu_X = 0$, $\mu_Y = 0$	
$r_{XX}(\tau) = \varphi_{XX}(\tau)$	$r_{XY}(\tau) = \varphi_{XY}(\tau)$
Autocovariance $\equiv$ autocorrelation	Cross-covariance $\equiv$ cross-correlation.

C.1.3.5 Temporal Averages and Ergodicity

In practical situations, it is not easy to calculate ensemble characteristics, which would require running simultaneously the same experiment on many copies of the plant under investigation. It is much easier to experiment with one *single* sample record $x_i(t)$ of the stochastic process, and to determine time averages calculated on this sample on a finite time horizon. We define therefore the

Temporal average of one sample $x_i(t)$:

$$\langle x_i(t)\rangle_T = \frac{1}{2T}\int_{-T}^{T} x_i(t)\,dt\,.$$

Although this time average is constant for a given sample, the set of corresponding values, if we could have taken them over all the samples, is a random variable, $\langle X(t)\rangle_T$, of which $\langle x_i(t)\rangle_T$ is one realization. From a practical point of view, it would be nice if we could estimate the ensemble mean of the stochastic process $X(t)$ from just one of its sample functions.

Now, for stationary processes, the ensemble average $\mu_X(t)$ is a constant μ_X. In this case, if

$$\lim_{T\to\infty} \mathrm{E}\left\{\langle X(t)\rangle_T\right\} = \mu_X \text{ and } \lim_{T\to\infty} \mathrm{Var}\left\{\langle X(t)\rangle_T\right\} = 0\,,$$

then the time-averaged mean and the ensemble mean are equal for $T \to \infty$. This property does not hold for *all* stationary processes, but only for a special (often encountered) class of random processes, the *ergodic* random processes.

A stationary stochastic process is said *ergodic*, when its time average and its ensemble mean are equal. Of course, in that case, the time average, at the limit $T \to \infty$, does not depend anymore on the sample $x_i(t)$ chosen.

C.1.3.6 Power Spectral Density (PSD)

Definition C.3. The power spectral density of a stationary random process $X(t)$ is the Fourier transform of the autocorrelation function $\varphi_{XX}(\tau)$ of this random process.

Thus, by adopting the symbolic writing $f(t) \leftrightarrow F(\omega) = \mathcal{F}\left[f(t)\right]$ to denote a pair composed of a function of time and its Fourier transform, we can write:

$$\varphi_{XX}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Phi_{XX}(\omega)\, e^{j\omega\tau} d\omega \;\leftrightarrow\; \Phi_{XX}(\omega) = \int_{-\infty}^{\infty} \varphi_{XX}(\tau)\, e^{-j\omega\tau} d\tau\,,$$

and, for two processes:

$$\varphi_{XY}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Phi_{XY}(\omega) e^{j\omega\tau} d\omega \;\leftrightarrow\; \Phi_{XY}(\omega) = \int_{-\infty}^{\infty} \varphi_{XY}(\tau) e^{-j\omega\tau} d\tau\,.$$

N.B.: $\Phi_{XX}(\omega)$ is a real function of ω because $\varphi_{XX}(\tau)$ is an even function of τ.

C.1.4 Discrete-time Random Processes

As in the continuous case, the first and second order ensemble characteristics of a discrete *stationary* stochastic process do not depend on the origin of time (see Table C.3). In particular, its ensemble mean is constant: $\mu_X(k)=\mu_X$.

Table C.3 Second order ensemble characteristics of one and two discrete random processes.

One process: $X(k)$	**Two processes:** $X(k)$ and $Y(k)$
1. General case (arbitrary processes):	
Autocovariance function:	Cross-covariance function:
$r_{XX}(j,k)=\mathrm{Cov}\{X(j),X(k)\}$ $=\mathrm{E}\{[X(j)-\mu_X(j)][X(k)-\mu_X(k)]\}$	$r_{XY}(j,k)=\mathrm{Cov}\{X(j),Y(k)\}$ $=\mathrm{E}\{[X(j)-\mu_X(j)][Y(k)-\mu_Y(k)]\}$
Autocorrelation function:	Cross-correlation function:
$\varphi_{XX}(j,k)=\mathrm{E}\{X(j)X(k)\}$	$\varphi_{XY}(j,k)=\mathrm{E}\{X(j)Y(k)\}$
2. *Stationary* processes: $j-k=m$	
Autocovariance function:	Cross-covariance function:
$r_{XX}(m)=\mathrm{Cov}\{X(k+m),X(k)\}$	$r_{XY}(m)=\mathrm{Cov}\{X(k+m),Y(k)\}$
Autocorrelation function:	Cross-correlation function:
$\varphi_{XX}(m)=\mathrm{E}\{X(k+m)X(k)\}$	$\varphi_{XY}(m)=\mathrm{E}\{X(k+m)Y(k)\}$
3. *Centered stationary* processes: $\mu_X=0\,;\mu_Y=0$	
$r_{XX}(m)=\varphi_{XX}(m)$	$r_{XY}(m)=\varphi_{XY}(m)$
4. Power spectral density:	
$\Phi_{XX}(\omega)=\sum_{k=-\infty}^{\infty}\varphi_{XX}(k)\,e^{-jk\omega}$ $\varphi_{XX}(k)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\Phi_{XX}(\omega)\,e^{jk\omega}d\omega$	$\Phi_{XY}(\omega)=\sum_{k=-\infty}^{\infty}\varphi_{XY}(k)\,e^{-jk\omega}$ $\varphi_{XY}(k)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\Phi_{XY}(\omega)\,e^{jk\omega}d\omega$

C.2 Random Vectors and Vectorial Random Processes

C.2.1 Definitions

The case of two random variables *X* and *Y*, seen in Sect. C.1.1.3, could have been considered as the situation of a two-dimensional random variable (X,Y).

The generalization to n dimensions follows then quite logically.

Random vector. An n-dimensional random variable $(X_1, X_2, \ldots, X_n)$ is often considered as a *random vector*

$$\mathbf{X} = \begin{pmatrix} X_1 & X_2 & \cdots & X_n \end{pmatrix}^{\mathrm{T}},$$

the components X_i of which are (scalar) random variables.

If the X_i are continuous random variables, $\mathbf{X}$ is a continuous random vector. This is the case considered here for the introduction of the new concepts, specific to the vector aspect of the situation. The corresponding definitions in the case of a discrete random vector will be deduced easily from that case by analogy.

Probability density. One calls probability density of the random vector $\mathbf{X}$ the function

$$p(x_1, x_2, \ldots, x_n) = p(\mathbf{x}),$$

such that

$$P\{\mathbf{x} \le \mathbf{X} < \mathbf{x} + d\mathbf{x}\} = p(\mathbf{x})\, d\mathbf{x},$$

and that

$$\int_{-\infty}^{\infty} p(\mathbf{x})\, d\mathbf{x} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \cdots dx_n = 1.$$

Mean or **expectation** of the random vector X:

$$\boldsymbol{\mu}_X = \mathrm{E}\{\mathbf{X}\} = \int_{-\infty}^{\infty} \mathbf{x}\, p(\mathbf{x})\, d\mathbf{x} = \begin{pmatrix} \mathrm{E}\{X_1\} & \cdots & \mathrm{E}\{X_n\} \end{pmatrix}^{\mathrm{T}}, \tag{C.8}$$

which is a *vector.*

Variance of the random vector **X**, and corresponding **covariance matrix**:

$$\begin{aligned} \boldsymbol{\Sigma}_{XX} &= \mathrm{Var}\{\mathbf{X}\} = \mathrm{E}\left\{(\mathbf{X} - \boldsymbol{\mu}_X)(\mathbf{X} - \boldsymbol{\mu}_X)^{\mathrm{T}}\right\} \\ &= \begin{pmatrix} \mathrm{Var}\{X_1\} & \mathrm{Cov}\{X_1, X_2\} & \cdots & \mathrm{Cov}\{X_1, X_n\} \\ \mathrm{Cov}\{X_2, X_1\} & \mathrm{Var}\{X_2\} & \cdots & \mathrm{Cov}\{X_2, X_n\} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}\{X_n, X_1\} & \cdots & \cdots & \mathrm{Var}\{X_n\} \end{pmatrix} \end{aligned} \tag{C.9}$$

Remark C.1. The variance of a random vector **X** is thus a matrix. Since its off-diagonal elements are covariances between the scalar coordinates of the vector, and even the diagonal ones by considering that $\mathrm{Var}\{X_i\} = \mathrm{Cov}\{X_i, X_i\}$, this matrix is usually called a *covariance matrix*. Since this might lead to confusion, some authors call it a *variance-covariance matrix*. For the sake of clarity, we will speak in this book of the *variance* of a vector, e.g. $\mathrm{Var}\{\mathbf{X}\}$, and of its *covariance matrix*, e.g. $\mathbf{\Sigma}_{XX}$, even though these two appellations cover the same mathematical concept.

Properties of the covariance matrix:

- $\mathbf{\Sigma}_{XX}$ is an $(n \times n)$, symmetric matrix:

$$\mathbf{\Sigma}_{XX} = \mathbf{\Sigma}_{XX}^{\mathrm{T}} \quad . \tag{C.10}$$

- When this matrix is diagonal, the (random) components X_i of the vector $\mathbf{X}$ are *uncorrelated*. If, further, the probability density is factorable, as

$$p(\mathbf{x}) = p(x_1)p(x_2)\ldots p(x_n) \,,$$

  the components X_i are *independent* (see Sect. C.1.1.3).

- The covariance matrix of a random vector is *always positive semidefinite*.

  Proof. Consider a random vector $\mathbf{X}$ of dimension n and its covariance matrix $\mathbf{\Sigma}_{XX} = \mathrm{Var}\{\mathbf{X}\}$. Since this matrix is symmetric, let us associate with it the following quadratic form, where $\mathbf{u}$ is a deterministic vector of same dimension as $\mathbf{X}$:

$$\mathbf{u}^{\mathrm{T}}\mathbf{\Sigma}_{XX}\mathbf{u} = \mathbf{u}^{\mathrm{T}}\mathrm{E}\left\{(\mathbf{X}-\boldsymbol{\mu}_X)(\mathbf{X}-\boldsymbol{\mu}_X)^{\mathrm{T}}\right\}\mathbf{u} = \mathrm{E}\left\{\mathbf{u}^{\mathrm{T}}(\mathbf{X}-\boldsymbol{\mu}_X)(\mathbf{X}-\boldsymbol{\mu}_X)^{\mathrm{T}}\mathbf{u}\right\}.$$

  Since $\mathbf{a}^{\mathrm{T}}\mathbf{b} = \mathbf{b}^{\mathrm{T}}\mathbf{a}$ for all vectors $\mathbf{a}$ and $\mathbf{b}$, the following holds:

$$\begin{aligned}\mathbf{u}^{\mathrm{T}}\mathbf{\Sigma}_{XX}\mathbf{u} &= \mathrm{E}\left\{\left[\mathbf{u}^{\mathrm{T}}(\mathbf{X}-\boldsymbol{\mu}_X)\right]^2\right\} \\ &= \mathrm{E}\left\{\left[u_1(X_1-\mu_{X_1}) + \ \ldots \ + u_n(X_n-\mu_{X_n})\right]^2\right\}.\end{aligned}$$

  Since the expectation of a square can only be positive or zero, it follows readily that $\mathbf{u}^{\mathrm{T}}\mathbf{\Sigma}_{XX}\mathbf{u} \geq 0$ whatever $\mathbf{u}$. The quadratic form can vanish when $\mathbf{u} = 0$ or when

$$u_1(X_1-\mu_{X_1}) + \ \ldots \ + u_n(X_n-\mu_{X_n}) = 0 \,,$$

i.e. when there exists a stochastic dependence between the components of the random vector **X** (see proof of the next property).

Therefore the quadratic form $\mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{XX}\mathbf{u}$, and thus also $\boldsymbol{\Sigma}_{XX}$, are *positive semidefinite.*

- The covariance matrix of a random vector whose components are statistically independent is always *positive definite.*

  Proof. Let us develop the previous quadratic form:

  $$\mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{XX}\mathbf{u} = \sum_i u_i^2 \operatorname{Var}\{X_i\} + 2\sum_{i,j} u_i u_j \operatorname{Cov}\{X_i, X_j\}.$$

  Since the X_i are statistically independent, $\operatorname{Cov}\{X_i, X_j\} = 0$, as seen in Sect C.1.1.3. Consequently, $\mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{u}$ can only vanish if $u_i = 0 \ \forall i$, thus if $\mathbf{u} = 0$.

Covariance of two random vectors **X** and **Y**:

$$\boldsymbol{\Sigma}_{XY} = \operatorname{Cov}\{\mathbf{X}, \mathbf{Y}\} = \operatorname{E}\left\{(\mathbf{X} - \boldsymbol{\mu}_X)(\mathbf{Y} - \boldsymbol{\mu}_Y)^{\mathrm{T}}\right\}. \tag{C.11}$$

Property:

$$\boldsymbol{\Sigma}_{XY} = \boldsymbol{\Sigma}_{YX}^{\mathrm{T}} . \tag{C.12}$$

C.2.2 Gaussian Random Vectors

Definition C.4. A random vector **X**, of dimension *n*, is *Gaussian* if it has a probability density of the form

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} \left|\boldsymbol{\Sigma}_{XX}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_X)^{\mathrm{T}} \boldsymbol{\Sigma}_{XX}^{-1} (\mathbf{x} - \boldsymbol{\mu}_X)}, \tag{C.13}$$

where $\boldsymbol{\mu}_X = \operatorname{E}\{\mathbf{X}\}$, $\boldsymbol{\Sigma}_{XX} = \operatorname{Var}\{\mathbf{X}\}$ and $\left|\boldsymbol{\Sigma}_{XX}\right|$ is the determinant of $\boldsymbol{\Sigma}_{XX}$.

In the Gaussian case, the probability distribution of the vector **X** is thus completely determined by $\boldsymbol{\mu}_X$ and $\boldsymbol{\Sigma}_{XX}$.

Property C.1. If $\boldsymbol{\Sigma}_{XX}$ is a diagonal matrix, the components of **X** are statistically independent, since $p(\mathbf{x})$ can then be factored as a product of *n* scalar distributions, according to (C.13). In other words, if the components of a Gaussian random vector are uncorrelated, they are statistically independent.

Property C.2. According to the properties of the variance (Sect. C.2.1), if the components of a Gaussian random vector $\mathbf{X}$ are uncorrelated, thus independent, its covariance matrix $\mathbf{\Sigma}_{XX}$ is positive definite.

Property C.3. Every linear combination of Gaussian random vectors is still a Gaussian random vector.

C.2.3 Vectorial Random Processes

The concept of scalar random processes, $X(t)$, is easily extended to that of a random process with n components, where the $X_i(t)$ are scalar stochastic processes:

$$\mathbf{X}(t) = \begin{pmatrix} X_1(t) & \cdots & X_n(t) \end{pmatrix}^{\mathrm{T}} .$$

$\mathbf{X}(t)$ is called a *vector-valued* or *vectorial stochastic process.*

One defines thus, similarly to the scalar case, the following notions, which will be presented briefly, first in the general case (non stationary case).

Mathematical expectation, or **mean** of $\mathbf{X}(t)$:

$$\mathrm{E}\{\mathbf{X}(t)\} = \boldsymbol{\mu}_X(t) .$$

Variance of $\mathbf{X}(t)$:

$$\mathrm{Var}\{\mathbf{X}(t)\} = \mathbf{\Sigma}_{XX}(t) = \mathrm{E}\left\{ [\mathbf{X}(t) - \boldsymbol{\mu}_X(t)][\mathbf{X}(t) - \boldsymbol{\mu}_X(t)]^{\mathrm{T}} \right\} .$$

Autocovariance of the process $\mathbf{X}(t)$:

$$\mathrm{Cov}\{\mathbf{X}(t_2), \mathbf{X}(t_1)\} = \mathbf{\Sigma}_{XX}(t_2, t_1) = \mathrm{E}\left\{ [\mathbf{X}(t_2) - \boldsymbol{\mu}_X(t_2)][\mathbf{X}(t_1) - \boldsymbol{\mu}_X(t_1)]^{\mathrm{T}} \right\} .$$

Autocorrelation of the process $\mathbf{X}(t)$:

$$\mathbf{\Lambda}_{XX}(t_2, t_1) = \mathrm{E}\left\{ \mathbf{X}(t_2)\mathbf{X}^{\mathrm{T}}(t_1) \right\} .$$

Remark C.2. To simplify writings, when there is no ambiguity, in particular when only one stochastic process $\mathbf{X}(t)$ is involved, the indices *XX* are often omitted.

In Table C.4 are collected the main definitions and properties of vectorial stochastic processes, both in continuous and discrete time. In this second case, an additional simplification of the writings consists in using indices, instead of parentheses, for the discrete time steps, each time there is no ambiguity.

Table C.4 Main concepts of continuous and discrete vectorial random processes.

Continuous case	Discrete case	
$\boldsymbol{\Sigma}(t)=\mathrm{Var}\left\{\mathbf{X}(t)\right\}$	$\boldsymbol{\Sigma}(k)=\boldsymbol{\Sigma}_k=\mathrm{Var}\left\{\mathbf{X}_k\right\}$	(C.14)
$\boldsymbol{\Sigma}(t_2,t_1)=\mathrm{Cov}\left\{\mathbf{X}(t_2),\mathbf{X}(t_1)\right\}$	$\boldsymbol{\Sigma}(j,k)=\boldsymbol{\Sigma}_{j,k}=\mathrm{Cov}\left\{\mathbf{X}_j,\mathbf{X}_k\right\}$	
1. Properties:		
$\boldsymbol{\Sigma}(t,t)=\mathrm{Var}\left\{\mathbf{X}(t)\right\}$	$\boldsymbol{\Sigma}(k,k)=\boldsymbol{\Sigma}_{k,\,k}=\mathrm{Var}\left\{\mathbf{X}_k\right\}$	(C.15)
$\boldsymbol{\Sigma}(t_1,t_2)=\boldsymbol{\Sigma}^{\mathrm{T}}(t_2,t_1)$	$\boldsymbol{\Sigma}_{k,\,j}=\boldsymbol{\Sigma}_{j,\,k}^{\mathrm{T}}$	
2. *Stationary* processes:		
$t_2-t_1=\tau$	$j-k=m$	
$\boldsymbol{\mu}_X(t)=\boldsymbol{\mu}_X;\,\boldsymbol{\Sigma}(t,t)=\boldsymbol{\Sigma}$	$\boldsymbol{\mu}_X(k)=\boldsymbol{\mu}_X;\,\boldsymbol{\Sigma}_{k,\,k}=\boldsymbol{\Sigma}$	
Autocovariance matrix:		
$\boldsymbol{\Sigma}(t_2,t_1)=\boldsymbol{\Sigma}(\tau)=\mathrm{Cov}\left\{\mathbf{X}(t+\tau),\mathbf{X}(t)\right\}$	$\boldsymbol{\Sigma}_{j,\,k}=\boldsymbol{\Sigma}_m=\mathrm{Cov}\left\{\mathbf{X}_{k+m},\mathbf{X}_k\right\}$	(C.16)
N.B.: the order is important:		
$\boldsymbol{\Sigma}(-\tau)=\boldsymbol{\Sigma}^{\mathrm{T}}(\tau)$	$\boldsymbol{\Sigma}_{-m}=\boldsymbol{\Sigma}_m^{\mathrm{T}}$	(C.17)
Autocorrelation matrix:		
$\boldsymbol{\Lambda}(t_2,t_1)=\boldsymbol{\Lambda}(\tau)=\mathrm{E}\left\{\mathbf{X}(t+\tau)\mathbf{X}^{\mathrm{T}}(t)\right\}$	$\boldsymbol{\Lambda}_{j,\,k}=\boldsymbol{\Lambda}_m=\mathrm{E}\left\{\mathbf{X}_{k+m}\,\mathbf{X}_k^{\mathrm{T}}\right\}$	
3. *Centered stationary* processes $(\boldsymbol{\mu}_X=0)$:		
$\boldsymbol{\Sigma}(\tau)=\boldsymbol{\Lambda}(\tau)$	$\boldsymbol{\Sigma}_m=\boldsymbol{\Lambda}_m$	
autocovariance matrix ≡ autocorrelation matrix		
4. Power spectral density:		
$\boldsymbol{\Phi}(\omega)=\int_{-\infty}^{\infty}\boldsymbol{\Lambda}(\tau)\,e^{-j\omega\tau}d\tau$	$\boldsymbol{\Phi}(\omega)=\sum_{k=-\infty}^{\infty}\boldsymbol{\Lambda}(k)\,e^{-jk\omega}$	(C.18)
$\boldsymbol{\Lambda}(\tau)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\boldsymbol{\Phi}(\omega)\,e^{j\omega\tau}d\omega$	$\boldsymbol{\Lambda}(k)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\boldsymbol{\Phi}(\omega)\,e^{jk\omega}d\omega$	(C.19)

Note that $\boldsymbol{\Lambda}(\tau)$, respectively $\boldsymbol{\Lambda}(k)$, can be replaced by $\boldsymbol{\Sigma}(\tau)$, respectively $\boldsymbol{\Sigma}(k)=\boldsymbol{\Sigma}_k$, in (C.18) and (C.19) in the centered case.

C.2.4 Gaussian Process (Scalar and Vectorial Cases)

A scalar stochastic process $X(t)$ is called Gaussian if, for any sequence of instants t_1, t_2, …, t_i, and whatever i, the random vector whose components are snapshots of the stochastic process taken at these instants, $\Xi = \left(X(t_1) \quad X(t_2) \quad \cdots \quad X(t_i) \right)^{\mathrm{T}}$, has a Gaussian distribution, as given by (C.13) with $\mathbf{X}$ replaced by Ξ.

The same definition applies of course also to a vectorial Gaussian stochastic process $\mathbf{X}(t)$: it is a stochastic process for which any sequence of random vectors $\mathbf{X}(t_1), \mathbf{X}(t_2), \ldots, \mathbf{X}(t_i)$ has a Gaussian joint distribution.

The particularity of a Gaussian process $\mathbf{X}(t)$ is that it is completely determined by the knowledge of its two first moments, $\mathrm{E}\left\{\mathbf{X}(t)\right\}$ and $\mathrm{Cov}\left\{\mathbf{X}(t_1), \mathbf{X}(t_2)\right\}$.

C.2.5 Discrete White Noise (Vectorial Case)

Definition C.5. A discrete white noise is a centered vectorial stochastic process $\mathbf{V}_k$, which consists of a sequence of uncorrelated random vectors. It is thus characterized by

$$\begin{cases} \mathrm{E}\left\{\mathbf{V}_k\right\} = 0, \ \forall k \\ \mathrm{Cov}\left\{\mathbf{V}_k, \mathbf{V}_j\right\} = \bar{\mathbf{Q}}_k \, \delta_{kj} \end{cases} \tag{C.20}$$

where $\bar{\mathbf{Q}}_k$ denotes a square, symmetric, positive semidefinite matrix (see Sect. C.2.1, Properties), which becomes a constant matrix $\bar{\mathbf{Q}}$ if the white noise is *stationary*.
If the white noise is in addition Gaussian, the random vectors $\mathbf{V}_k$ are independent, and consequently $\bar{\mathbf{Q}}_k$ is positive definite (see Sect. C.2.2, Property C.2).

C.2.6 Continuous White Noise (Vectorial Case)

Definition C.6. A continuous white noise is a centered vectorial stochastic process $\mathbf{V}(t)$, which consists of uncorrelated random vectors $\mathbf{V}(t)$, $\mathbf{V}(t')$, whatever $t \neq t'$. It is thus characterized by

$$\begin{cases} \mathrm{E}\left\{\mathbf{V}(t)\right\} = 0, \quad \forall t \\ \mathrm{Cov}\left\{\mathbf{V}(t), \mathbf{V}(t')\right\} = \bar{\mathbf{Q}}(t)\, \delta(t - t') \end{cases} \tag{C.21}$$

where $\bar{\mathbf{Q}}(t)$ denotes a square, symmetric, positive semidefinite matrix (see Sect. C.2.1, properties), which becomes a constant matrix $\bar{\mathbf{Q}}$ if the white noise is *stationary*.

If the white noise is in addition Gaussian, the random vectors $\mathbf{V}(t)$ and $\mathbf{V}(t')$, with $t \neq t'$, are independent, and consequently $\bar{\mathbf{Q}}(t)$ is positive definite (see Sect. C.2.2, Property C.2).

Spectral property. With (C.18) we deduce from (C.21) that the power spectral density of stationary continuous white noise is constant, and equal to $\bar{\mathbf{Q}}$.

Remark C.3. The above definitions C.5 and C.6 hold of course also in the scalar situation.

D Simulation Diagrams

D.1 Simulink® Universal Simulation Diagrams

The diagrams in Fig. D.1 and Fig. D.2 for continuous-time and discrete-time simulations are of a universal nature, in the sense that they are adequate for all the exercises proposed in the book. They comprise a state-feedback controller, an observer in its generalized form, a gain compensation (feedforward) matrix and error integrators. The parameters the user will have to adjust are listed in Table D.1. In the two diagrams, the **C** parameter of the *State-Space* block simulating the plant is an identity matrix, of suitable dimensions, in order to give access to the full state vector at the output of this block. The measurement matrix $\mathbf{C}_m$ which builds the real plant output is introduced separately in the diagram. The selection matrix $\mathbf{S}_i$ is used for the feedback with integral action; the matrix $\mathbf{S}_c$ is used in conjunction with the gain compensation (or anticipation) term **M** (for more details, see Chap. 1, Sect. 1.8.3 and 1.1.3).

Manual switches enable the user to disable unused functions, such as e.g. the observer (switch S_5) in the case that the simulation of a direct state feedback is desired, this state being assumed entirely accessible. If no observer has been calculated yet at this point, the corresponding block parameters are set to zero by the program.

The S_1 switch gives the choice between a step reference and a pulse generator. It is to the user to set, in the first case the amplitude parameter (*final value*) $\mathbf{y}_r$ and the application time (*step time*) *start_reference* of the step function, in the second case the amplitude, period and pulse width of the generated signal, by respecting the dimension of the involved quantities (scalars of vectors). For more details about the dimensions of the output signals of these two blocks, the reader is invited to consult the on-line documentation of MATLAB®. Switches S_3 and S_4 enable applying constant load and measurement perturbations. It is again to the user to make appropriate choices for the amplitudes *p_load* , respectively *p_measurements*, and the application times *start_p_load*, respectively *start_p_measurements*, of these perturbations, with the same remark as above as to their dimensions. In the case of the discrete-time diagram, the switches S_6 and

E. Ostertag, *Mono- and Multivariable Control and Estimation,*
DOI 10.1007/978-3-642-13734-1_10, © Springer-Verlag Berlin Heidelberg 2011

S_7 enable applying random disturbances, which are useful for studying noisy systems and optimal filtering.

The error integrators are aimed at simulating state-feedback controls *with integral action*, as described in Sect. 1.8. They can be disabled in this diagram, when not used, by choosing $\mathbf{L}_1 = \mathbf{L}$ and $\mathbf{L}_2 = 0$.

The return to equilibrium without any input signal and from a non vanishing initial state $\mathbf{x}_0$ of the plant and/or an initial state $\mathbf{z}_0$ of the observer can be studied by swapping the switch S_2 with of course in this case $\mathbf{L}_2 = 0$.

In the various diagrams the user will have to adjust the number of channels (*Number of axes*) of the oscilloscopes, since this parameter can be introduced only numerically. After having checked the box *Save data to workspace*, he will also need to fill in the field *Variable name* which appears under the tab *Data history* with the corresponding name used in the subroutine *Simulation_Scopes.m* included in *mmce.zip*.

Although not essential, zero-order holds are inserted in front of the oscilloscopes in the discrete-time diagram, in order to enforce the staircase aspect of the oscillograms, since otherwise the oscilloscopes would interpolate the plots between measurement points.

Table D.1 Parameters of the simulation blocks.

<table>
<tr><td><i>Continuous state-space model (plant):</i>
A: A B: [B E] C: eye(n)
D: zeros(n, p+re)
Initial conditions : [x0]</td><td><i>Discrete state-space model (plant):</i>
A: Phi B: [Gamma E] C: eye(n)
D: zeros(n, p+re)
Initial conditions: [x0]; Sample time: Ts</td></tr>
<tr><td>Generalized observer:
A: F B: [J G] C: H
D: [zeros(n,p) K]
Initial conditions: [z0]</td><td>Generalized observer:
A: F B: [J G] C: H
D: [zeros(n,p) K]
Initial conditions: [z0]; Sample time: Ts</td></tr>
<tr><td colspan="2"><i>Step</i>: Step time: start_reference; Initial value: 0; Final value: yr; Sample time: 0
<i>Pulse generator</i>: Amplitude: yr; Period: simulation_horizon/2.5; Pulse width: 50%
<i>Load disturbance</i>: Step time: start_p_load; Initial value: 0; Final value: p_load (p_load*Ts in the discrete case)
<i>Measurement disturbance</i>: Step time: start_p_measurements; Initial value: 0; Final value: p_measurements</td></tr>
<tr><td colspan="2"><i>Simulation / Configuration Parameters:</i>
Solver stop time: simulation_horizon
Solver options: Variable-step (continuous); Fixed-step; Fixed step size: Ts (discrete)</td></tr>
</table>

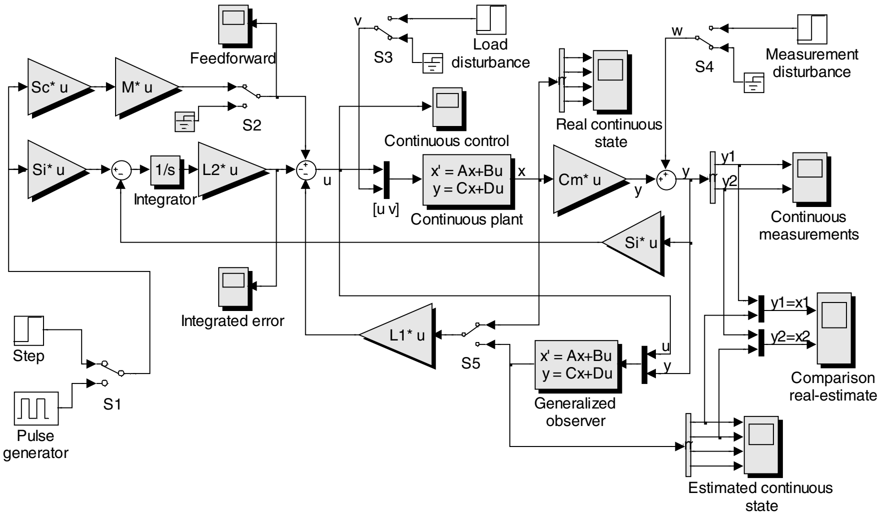


Fig. D.1 Universal simulation diagram for continuous-time state-feedback control laws and observers (*Universal_continuous_simulation.mdl*).

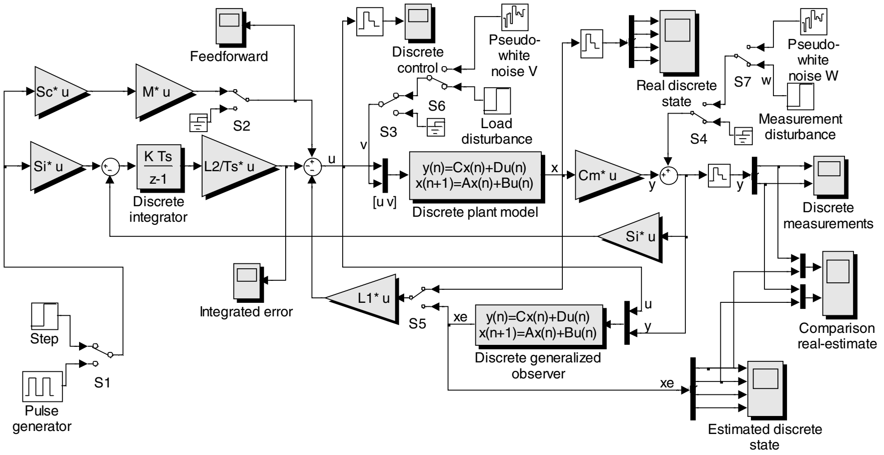


Fig. D.2 Universal simulation diagram for discrete-time state-feedback control laws and observers (*Universal_discrete_simulation.mdl*).

D.2 Specific Simulation Diagram for the Inverted Pendulum

The third simulation diagram is on the contrary a specific discrete-time diagram, which serves only for the simulation of the inverted pendulum and takes into account its nonlinearity and the compensation of it. It is used exclusively for the simulations of Exercise 2 of Chapter 2.

A *Coulomb & Viscous Friction* block is used to model the dry friction. The *Coefficient of viscous friction* of this block is set to zero. The switch S_2 activates of disables this nonlinearity.

The observer, of reduced order, is represented in this diagram in details by means of the various matrices involved in its generalized representation. The switch S_3 commutes the dry friction compensation mode from *fixed threshold* to *estimation by observer*. The switch S_4 enables suppressing any compensation.

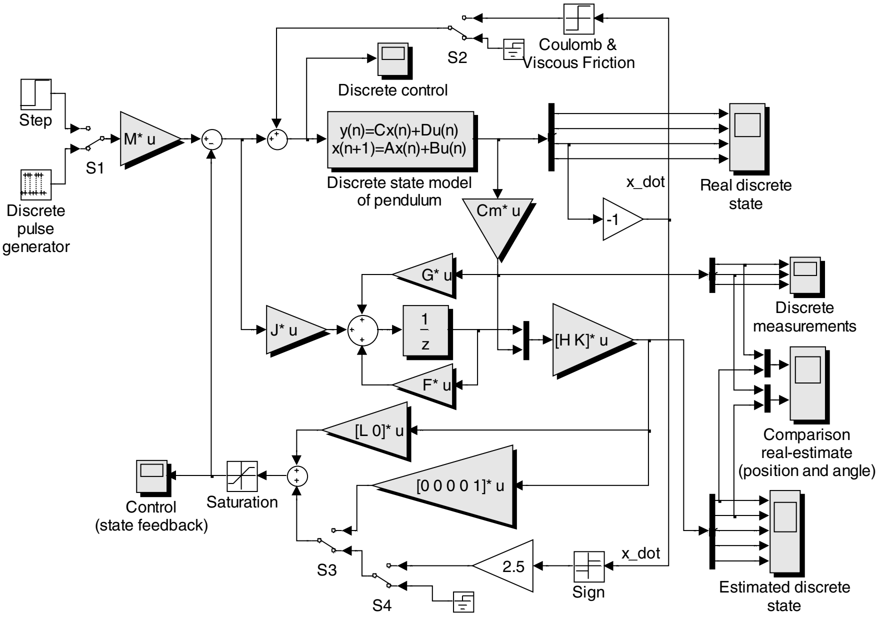


Fig. D.3 Simulation diagram for the inverted pendulum, with disturbance observer (*Pendulum_and_frictions.mdl*).

References

[Ack72] J. Ackermann. “Der Entwurf linearer Regelungssyteme im Zustandsraum”, *Regelungstechnik*, **20**, pp. 297-300, 1972.

[Ala05] D. Alazard. *Introduction to Kalman Filtering*. European Masters Course in Aeronautics and Space Technology, SUPAERO, 2005.

[AlSa04] P. Albertos, A. Sala. *Multivariable Control Systems. An Engineering Approach*, Springer, 2004.

[Ami92] Amira™ GmbH. *Laboratory Setup Inverted Pendulum LIP100*, manuel technique, 1992

[ANAL06] T. Alamo, J.E. Normey-Rico, M.R. Arahal, D. Limon, E. Camacho. “Introducing linear matrix inequalities in a control course”, *Proc. Advances in Contr. Educ., ACE’06,* Madrid, 2006.

[AnMo89] B.O. Anderson, J.B. Moore. *Optimal Control: Linear Quadratic Methods*, Prentice Hall, 1989.

[ArBC93] D. Arzelier, J. Bernussou, G. Garcia. “Pole assignment of linear uncertain systems in a sector via a Lyapunov-type approach”, *IEEE Trans. Aut. Control,* **38**, pp. 1128-1131, 1993.

[AsWi97] K.J. Åström, B. Wittenmark. *Computer Controlled Systems*. *Theory and Design*, 3rd ed., Prentice-Hall, 1997.

[BaBO88] N. Bakri, N. Becker, E. Ostertag. “Anwendung von Kontroll-Störgrößenbeobachtern zur Regelung und zur Kompensation trockener Reibung”, *Automatisierungstechnik* ***at***, **36**, pp. 50-54, 1988.

[BBFE93] S. Boyd, V. Balakrishnan, E. Feron, L. El Ghaoui. “Control system analysis and synthesis via linear matrix inequalities”, *Proc. Amer. Contr. Conf.*, San Francisco, pp. 2147-2154, 1993.

[BEFB94] S.P. Boyd, L. El Ghaoui, E. Feron, V. Balakrishnan. *Linear Matrix Inequalities in Systems and Control Theory*, vol. 15 of *Studies in Applied Mathematics*, SIAM, Philadelphia, 1994.

[BeOs87] N. Becker, E. Ostertag. “Zur Berechnung der Zustandsrückführmatrix für Strecken mit mehreren Eingangsgrößen”, *Automatisierungstechnik* ***at*****,** **5**, pp. 214-215, 1987.

[BoDu96] P. Boucher, D. Dumur. *La commande prédictive*, éd. Technip, 1996.

E. Ostertag, *Mono- and Multivariable Control and Estimation,*
DOI 10.1007/978-3-642-13734-1, © Springer-Verlag Berlin Heidelberg 2011

[BoYo68] J.J. Bongiorno, D.C. Youla. "On observers in multivariable control systems", *Int. J. Control*, **8**, n°3, pp. 221-243, 1968.

[BrSi75a] K. Brammer, G. Siffling. *Stochastische Grundlagen des Kalman-Bucy-Filters: Wahrscheilnlichkeitsrechnung und Zufallsprozesse* Oldenburg, 1975.

[BrSi75b] K. Brammer, G. Siffling. *Kalman-Bucy-Filter: Deterministische Beobachtung und stochastische Filterung,* Oldenburg, 1975.

[CAJZ01] A. Crosnier, G. Abba, B. Jouvencel, R. Zapata. *Ingénierie de la commande des systèmes*, Ellipses, Coll. Technosup, 2001.

[CeBa84] G. Celentano, A. Balestrino. "New techniques for the design of observers", *IEEE Trans. Aut. Cont.*, **AC-29**, n° 11, pp. 1021-1025, 1984.

[ChGa96] M. Chilali, P. Gahinet, "$\mathcal{H}_\infty$ Design with pole placement constraints: An LMI approach," *IEEE Trans. Aut. Control,* **41**, pp. 358-367, 1996.

[ClMo87] D.W. Clarke, C. Mohtadi. "Properties of Generalized Predictive Control", *Proc. 10th World Cong. Aut. Cont.*, IFAC'87, Munich, **10**, pp. 63, 1987.

[Doy78] J.C. Doyle. "Guaranteed margins in LQG regulators", *IEEE Trans. Aut. Control,* **AC-23,** n°4, pp. 664-665, 1978.

[Duc01] G. Duc. *Commande par variables d'état des systèmes linéaires*, lecture notes, Supélec n° 03145, 2001.

[Duc02a] G. Duc. *Analyse des systèmes linéaires*, sous la direction de Ph. De Larminat, Hermès, chap. 6, « Le formalisme de Kalman pour la stabilisation et l'estimation d'état », pp. 171-201, 2002.

[Duc02b] G. Duc. *Introduction aux LMI*, lecture notes, Supélec, 2002.

[Duc04] G. Duc. *La commande optimale des systèmes dynamiques*, sous la direction de H. Abou-Kandil, Hermès, chap. 3, « Systèmes linéaires », pp. 117-167, 2004.

[FaWo67] P.L. Falb, W.A. Wolovich. "Decoupling in the design and synthesis of multivariable control systems", *IEEE Trans. Aut. Cont.*, **12**, pp. 651-659, 1967.

[Fee82] Feedback Instruments™ Ltd. *Process Trainer PT326*, manuel technique, 1982.

[Föl90] O. Föllinger. *Regelungstechnik*, Hüthig, Heidelberg, 6th ed., Chap. 13, pp. 464-526, 1990.

[FöRo88] O. Föllinger, unter Mitwirkung von G. Roppenecker. *Optimierung dynamischer Systeme – Eine Einführung für Ingenieure*, Oldenburg, 1988.

[FrPE94] G.F. Franklin, J.D. Powell, A. Emami-Naeini. *Feedback Control of Dynamic Systems*, 3rd ed. Addison-Wesley, 1994.

[FrPW97] G.F. Franklin, J.D. Powell, M. Workman. *Digital Control of Dynamic Systems*, 3rd ed. Addison-Wesley, 1997.

[Gee07] H.P. Geering. *Optimal Control with Engineering Applications*, Springer, 2007.

[GoOs03] E. Godoy, E. Ostertag. *Commande numérique des systèmes : approches fréquentielle et polynomiale*, Ellipses, pp. 32-35, and. 165-222, 2003.

[Gra03] Y. Granjon. *Automatique: systèmes linéaires, non linéaires, à temps continu, à temps discret, représentation d'état*, Dunod, 2003.

[GrBo08] M. Grant, S. Boyd. Graph implementations for nonsmooth convex programs, *Recent Advances in Learning and Control (a tribute to M. Vidyasagar)*, V. Blondel, S. Boyd, and H. Kimura, ed., p. 95-110, *Lecture Notes in Control and Information Sciences*, Springer, 2008. http://stanford.edu/~boyd/graph_dcp.html.

[GrBo10] M. Grant, S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21. http://cvxr.com/cvx, May 2010.

[Kay06] S.M. Kay. *Intuitive Probability and Random Processes using MATLAB*, Springer, 2006.

[Kal60] R.E. Kalman. "A new approach to linear filtering and prediction problems", *J. Basic Eng., ASME Trans.*, **82D**, n°1, pp. 35-45, 1960.

[Lar93] P. de Larminat. *Automatique : commande des systèmes linéaires*, Hermès, 1993.

[LaTh77] P. de Larminat, Y. Thomas. *Automatique des systèmes linéaires*, Tome 2 : *Identification*, Hermès, chap. 5, pp. 137-177, 1977.

[Lev96] W.S. Levine. *The Control Handbook*. A contributed book from over 170 authors, CRC Press, 1996.

[LiKM94] T. Livet, F. Kubica, J.F. Magni. "Performance improvement by feed-forward: application to civil aircraft control design", *IEEE*, pp. 341-346, 1994.

[LKMA94] T. Livet, F. Kubica, J.F. Magni, L. Antonel. "Non-interactive control by eigenstructure assignment and feedforward", *AIAA Conf. Guidance, Navigation and Control*, Phoenix, USA, 1994

[Löf04] J. Löfberg, "YALMIP: A toolbox for modelling and optimization in MATLAB," in *Proc. CACSD Conference*, Taipei, Taiwan, 2004. [Online] available (2010): http://users.isy.liu.se/johanl/yalmip/

[Lue64] D.G. Luenberger. "Observing the state of a linear system", *IEEE Trans. Mil. Electr.*, **8**, pp. 74-80, 1964.

[Lue66] D.G. Luenberger. "Observers for multivariable systems", *IEEE Trans. Aut. Cont.*, **11**, pp. 190-197, 1966.

[Lue71] D.G. Luenberger. "An introduction to observers", *IEEE Trans. Aut. Cont.*, **16**, pp. 596-602, 1971.

[Lun06] J. Lunze. *Regelungstechnik 1 : Systemtheoretische Grundlagen, Analyse und Entwurf einschleifiger Regelungen*, 5. Auflage, Springer, 2006

[Mey01] C.D. Meyer. *Matrix Analysis and Applied Linear Algebra*, SIAM, Philadelphia, 2001

[MoGB81] J.B. Moore, D. Gangsaas, J. Blight. "Performance and robustness trades in LQG regulator designs", *Proc. 20th IEEE Conf. on Dec. and Contr.,* San Diego, USA, pp. 1191-1199, 1981.

[MoMa98] M. Mokhtari, M. Marie. *Applications de MATLAB® 5 et SIMULINK® 2*, Springer, 1998.

[MuBa64] F.J. Mullin, J. DeBarbeyrac. "Linear digital control", *J. Basic Eng., ASME Trans.*, **86**, pp. 61-66, 1964.

[Oga02] K. Ogata. *Modern Control Engineering*, 4th ed. Prentice Hall, 2002.

[Ost04] E. Ostertag. *Systèmes et asservissements continus : modélisation, analyse, synthèse des lois de commande*, Ellipses, 2004.

[Rib04] M.I. Ribeiro. "Kalman and Extended Kalman Filters: Concept, Derivation and Properties", 2004. [Online]: mir@isr.ist.utl.pt .

[Rop81] G. Roppenecker. "Polvorgabe durch Zustandsrückführung", *Regelungstechnik*, **29**, pp. 228-233, 1981.

[Rop82] G. Roppenecker. "Äquivalenz zweier Darstellungsformen des Polvorgabereglers bei Eingrößensystemen", *Regelungstechnik*, **30**, pp. 212-213, 1982.

[Rop86] G. Roppenecker. "On parametric state feedback design", *Int. J. Control*, **43**, pp. 793-804, 1986.

[Rop90] G. Roppenecker. *Zeitbereichsentwurf linearer Regelungen*, Oldenburg, 1990.

[Ros62] H.H. Rosenbrock. "Distinctive problems of process control", *Chemical Engineering Progress*, **58**, pp.43-50, 1962.

[ScGC97] C. Scherer, P. Gahinet, M. Chilali. "Multiobjective output-feedback control via LMI optimization", *IEEE Trans. Aut. Control*, **42**, pp. 896-911, 1997.

[ScWe06] C. Scherer, S. Weiland. *Linear Matrix Inequalities in Control*, Report, Delft University of Technology, 2006. Available online.

[ShBr88] K.Sam Shanmugan, A.M. Breipohl. *Random Signals: Detection, Estimation and Data Analysis*, Wiley, 1988.

[Sim06] D.J. Simon. *Optimal State Estimation: Kalman, H_infinity and nonlinear approaches*, Wiley, 2006

[Stu99] J.F. Sturm, "Using SeDuMi 1.02: A Matlab toolbox for optimization over symmetric cones," *Optimization Methods and Software*, vol. 11-12, pp. 625-653, 1999. [Online],
version 1.1R3 (2006): http://sedumi.mcmaster.ca/.
version 1.21 (2009) and 1.3 (2010): http://sedumi.ie.lehigh.edu/.

[Won67] W.M. Wonham. "On pole assignment in multi-input controllable linear systems", *IEEE Trans. Aut. Control*, **12**, pp. 600-665, 1967.

[ZhDG96] K. Zhou, J.C. Doyle, K. Glover. *Robust and Optimal Control*, Prentice Hall, Chap. 13, pp. 327-341, and Chap. 7, 1996.

Index

E. Ostertag, *Mono- and Multivariable Control and Estimation,*
DOI 10.1007/978-3-642-13734-1, © Springer-Verlag Berlin Heidelberg 2011